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

Version: 4th Edition
Copyright date: 2003
Grade or Age Range: 5 through 21 years
Author: Eleanor Semel, Ed.D. ; Elisabeth H. Wiig, Ph.D.; and Wayne A. Secord, Ph.D.
Publisher: Pearson

1. PURPOSE

The CELF-4 is designed to assess the presence of a language disorder or delay using a comprehensive and flexible assessment approach. Subtests were designed in correspondence with educational mandates with regards to (a) eligibility for services, (b) identification of strengths and weaknesses, and (c) performance within tasks related to the standard educational curriculum. The CELF-4 includes a four-level assessment process in which the presence of a language disorder can be determined by calculating a Core Language score using only four subtests. Content areas include: morphology and syntax, semantics, pragmatics, and phonological awareness.

2. DESCRIPTION

Subtest	Age Range	Purpose	Format
Concepts and Following Directions[1,2]	5-12	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.
Word Structure[1]	5-8	To determine the student's use of morphological rules.	The student completes a sentence that is orally presented by the administrator in reference to a visual stimulus.

Formulated Sentences[1,2, 3]	5-21	To measure a student’s ability to formulate grammatically and semantically correct sentences.	Following an orally presented target word from the administrator, the student generates a sentence in reference to a visual stimulus.
Recalling Sentences[1,2, 3]	5-21	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.
Word Classes 1	5-7	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.
Word Classes 2 [2, 3]	8-21	To measure an individual’s ability to comprehend and explain relationships between orally presented target words.	Given 3-4 words, the student selects two words that go together and explains their relationship.
Word Definitions [3]	10-21	To measure a student’s ability to infer word meanings based on class relationships and shared meanings.	Following an orally presented target word that is used in a sentence, the student defines the word.
Sentence Structure	5-8	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.
Expressive Vocabulary	5-9	To measures a student’s ability to use referential naming.	The student identifies an object, person, or action presented by the administrator.
Understanding Spoken Paragraphs	9-21	To measure an individual’s ability to comprehend a story by answering factual questions and making inferences based on story material.	Following presentation of an orally presented story, the student answers questions.
Sentence Assembly	13-21	To measures an individual’s ability to formulate syntactically and semantically correct sentences.	Following presentation of visually and verbally presented words, the student formulates two sentences.

Semantic Relationships	13-21	To measure a student's ability to comprehend sentences that
		(a) Make comparisons;
		(b) Use location;
		(c) Use time relationships;
		(d) Use serial order;
		(e) Use passive voice
			Following an orally presented sentence, the student identifies the two correct choices from a field of four.

Phonological Awareness	5-12	To measure a student's acquisition of sound structure and ability to manipulate sound through: (a) Rhyme; (b) Syllable, phoneme, and sentence segmentation; (c) Syllable and phoneme blending; (d) Syllable identification; (e) Phoneme manipulation and identification.	Comprised of 17 tasks of varying directives.
Rapid Automatic Naming	5-21	To measure the student's ability to produced automatic speech.	The student is timed during naming of color, shapes, and color-shape combinations.
Word Associations	5-21	To measure the student's ability to recall objects from a semantic category within a fixed time limit.	The student lists objects belonging to a semantic category within one minute.
Number Repetition 1	5-16	To measure the student's working memory.	The client repeats a series of digits in the exact order presented by the administrator. Following, the student repeats digits in reverse order of an orally presented string of numbers.
Number Repetition 2	17-21	To measure the student's working memory.	The client repeats a series of digits in the exact order presented by the administrator. Following, the student repeats digits in reverse order of an orally presented string of numbers.

Familiar Sequences 1	5-16	To measure the student's ability to retrieve common information.	The student recites common information (e.g. days of the week, counting backwards, etc.) while being timed.
Familiar Sequences 2	17-21	To measure the student's ability to retrieve common information.	The student recites common information (e.g. days of the week, counting backwards, etc.) while being timed.

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

3. STANDARIZATION SAMPLE

The standardization sample for the CELF-4 used data collected in 2002 and was comprised of a sample of over 4,500 individuals from 47 states aged 5 through 21 years. Inclusion into the standardization sample required completion of the test in a standard manner (e.g. no sign language was permitted). The standardization sample was stratified by demographic factors including age, gender, race, parental education level, and geographic location as compared to the 2000 national census. English was identified as the primary language for all subjects; however, approximately 15% of the sample population was from homes that spoke a language other than English. Approximately 9.5% of the sample reported receiving special related services at the time of testing, including 2.6% for gifted and talented, 2.8% for learning disabilities, 2% for intellectual disabilities, and <1% each for hearing impairment, visual impairment, and autistic disorders. Additionally, 7% of the students in the sample were receiving speech and language services for disorders including articulation, language, auditory habilitation, voice disorders, fluency, phonology, and pragmatics. The manual did not state whether or not the students who were previously identified as having a disability were accurately identified by the CELF-4. According to Peña, Spaulding, & Plante (2006), inclusion of individuals with disabilities in the normative sample can negatively impact the test's discriminant accuracy, or its ability to differentiate between typically developing and disordered children. Specifically, when individuals with disabilities are included in the normative sample, the mean scores are lowered. As a result, children will only be identified as having a disability with an even *lower* score. Thus, children with mild disabilities will not be identified, compromising the sensitivity of the test.

4. VALIDITY

Content - Content validity is how representative the test items are of the content that is being assessed (Paul, 2007). Content validity was analyzed using a literature review and feedback from focus groups. A nationwide pilot study was conducted in August 2000 using 86 students aged 5-21. The sample consisted of 67% Caucasian students, 16% Hispanic students, 10% African American students, and 7% Asian students. Results from the pilot study resulted in elimination and revision of subtests to meet the goals of the CELF-4 revision. Additionally, a nation-wide tryout test using 14 CELF-4 subtests was conducted by 338 speech-language pathologists. The sample population for this study was comprised of 2,259 typically developing students aged 5-21, who spoke English as their primary language. In addition, a separate clinical study was conducted using 513 students with language disorders, who also spoke English as their primary language. Specific results were not provided; however, responses obtained during the tryout test and clinical studies were analyzed by a scoring panel. In addition, no information was provided regarding how it was determined that the students spoke "English as their primary language" (See p. 203-204 of the manual for further information). 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.

Finally, an expert panel comprised of nine speech-language pathologists evaluated the content validity of the CELF-4. Information regarding their level of expertise was not provided, and expert knowledge of a variety of dialects requires an enormous and sophisticated knowledge base. In some cases, the intricacies of dialectal variations are so small that even highly educated linguists find it difficult to determine differences between cultures. As specific information regarding the background and training of the "expert panel" was not provided, one cannot be confident that the items in this test are completely free of bias. Further, the diagnostic status of the pilot study participants was not provided in the manual and thus we cannot be certain that all students in that sample were typically developing; therefore, the content validity is considered insufficient.

Concurrent - Concurrent validity is the extent to which a test agrees with other valid tests of the same measure (Paul, 2007). According to McCauley & Swisher (1984) concurrent validity can be assessed using indirect estimates involving comparisons amongst other tests designed to measure *similar behaviors*. If both test batteries result in similar scores, the tests "are assumed to be measuring the same thing" (McCauley & Swisher, 1984, p. 35). Concurrent validity was measured by comparing the CELF-4 to the CELF-3 using two studies. The first study used a sample of 158 typically developing children ages 6 through 13. Correlation coefficients for this study were .84 for Core Language Score, and .79 for both Receptive and Expressive Language scores. A second study was conducted using a group of 57 individuals who were diagnosed with language disorders. No information regarding the age of the students was reported. The correlation coefficients for this study were .8 for Core Language score, .73 for Receptive Language score, and .71 for Expressive Language score.

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). In order to examine the validity of a comparison measure, it is important to consider its discriminant accuracy, or its sensitivity and specificity. The sensitivity for the CELF-3 was reported at 78.7%, meaning 21.3% of children who had a language disorder were *not identified* as having a disorder by the CELF-3 (Semel, Wiig, & Secord, 1995). Specificity was reported as 92.6%, meaning 7.4% of typically developing children were falsely identified as having a language disorder. Although the specificity is considered “good” by the standard in the field, the sensitivity is considered “unacceptable” (Plante & Vance, 1994), and thus, the CELF-3 is not a valid instrument to use as a comparison test. Further, as the study of typically developing students only compared children aged 6-13, and age of the children with language disorders was not reported, neither study covers the age range of the CELF-4 (5-21), making it an insufficient comparison. The concurrent validity is therefore invalid due to the poor diagnostic accuracy of the comparison test and the insufficient assessment of the full age range of the CELF-4.

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

Reference Standard

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

The reference standard used to identify children as having a language disorder (part of the sensitivity group) was a score below 1.5 SD on a standardized test of language skills. The clinical study group included 225 children, recruited by speech language pathologists across the United States, aged 6-17 years. It is important to note that this does not include the entire age range of the CELF-4, and thus is not representative of the test population. Standardized tests that were used to identify the children in the reference standard include the CELF-3 (51.11%), CELF-3 and another diagnostic instrument (6.66%), Test of Language Development (TOLD) (16.44%), PLS-3 (8.44%), Oral and Written Language

Scales (OWLS) (8%), Comprehensive Assessment of Spoken Language (3.56%) and "other" (5.78%). Over half of the students were identified using the CELF-3 and the PLS-3, both of which have been identified as instruments with unacceptable levels of sensitivity (Plante & Vance, 1994). It is important to note that arbitrary cut off scores on standardized language tests often do not accurately discriminate between typically developing children and children with a language disorder (Spaulding, Plante, & Farinella, 2006). Thus, the true diagnostic status is unknown. The reference standard is insufficient because it does not include the entire age range of the CELF-4 and students were included in the reference standard based on previous test performance on potentially invalid measures with an arbitrary cut off score.

The reference standard used to identify the *specificity* group was a control group of 225 typically developing students, selected from the standardization sample, who were matched to the sensitivity group based on age, gender, race, and level of parental education. It was not defined how the children used for the specificity group were identified as typically developing. The reference standard for the specificity group is insufficient for several reasons. First, it does not include students from the entire age range of the index and therefore is not representative of the population. Further, as the students were chosen from the standardization sample, their "true" diagnostic status cannot be determined; 9.5% of the standardization sample was receiving special services and 7% was receiving SLP services. According to Dollaghan (2007) performance on the reference standard cannot be assumed. As the same reference standard (a score above 1.5 SD below the mean on a standardized test) for the sensitivity group was not applied to the specificity group, one cannot be certain of the diagnostic status of the control group.

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 the limits to be identified as having a language disorder on this assessment. The CELF-4 reports sensitivity measures to be 1.00 for cut-off scores of -1 and -1.5 standard deviations (SD) below the mean and 0.87 for cut-off scores of -2 SD below the mean. According to Plante & Vance (1994), validity measures above .9 are good, measures between .8 and .89 are fair, and measures below .8 are unacceptable. Therefore, the sensitivity reported in the manual of the CELF-4 is considered "fair" to "good." However, it is important to consider the implications of these measures. A sensitivity of .87 means that 13/100 children who have a language disorder will not be identified as such by the CELF-4, 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-4 reports specificity measures to be 0.82 at -1 SD below the mean, 0.89 at -1.5 SD below the mean, and 0.96 at -2 SD below the mean. Although these measures are considered "fair" to "good" (Plante & Vance, 1994), it is important to consider the implications. A specificity of .82 means that 18/100 typically developing children will be identified as having a language disorder and may be unnecessarily referred for special education services. Further, as the reference standard (score below 1.5 SD on a standardized language assessment) was not applied to the specificity group, it is important to consider a possible spectrum bias (Dollaghan & Horner, 2011). Since the children included in the specificity sample were chosen from the standardization sample, and were not administered the same reference standard, their diagnostic status cannot be determined for certain. As mentioned above, about 9.5% of the standardization sample was receiving special related services and 7% of the sample were receiving speech and language services. Therefore, the children in the specificity sample cannot be guaranteed to be free of a disorder.

Despite the "fair" to "good" sensitivity and specificity measures, the construct validity of the CELF-4 is weak as 66% of the 225 individuals in the sensitivity group (i.e. considered to have a speech and language disability) were included based on their scores on the CELF-3 or PLS-3. Both the PLS-3 and CELF-3 have "unacceptable" levels of sensitivity and specificity. The sensitivity of the PLS-3 ranged from 36%-61% for 3-5 year olds (Zimmerman, Steiner, & Pond, 1991, as cited in Crowley, 2010). The sensitivity and specificity for the CELF-3 were reported as 78.7% (but 57% when considering children with language disorders alone) and 92.6% respectively (Ballantyne, Spilkin, & Trauner, 2007, as cited in Crowley, 2010). The CELF-4 cannot claim that it is valid because its scores match those of an invalid and inaccurate test. Further, the children used to identify the specificity were chosen from the standardization sample and as noted above, children with disabilities were included in that sample, making the reference standard insufficient.

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. As the LR+ and LR- are calculated using the sensitivity and specificity and these measures have been shown to be insufficient as compared to the standards in the field, likelihood ratios are not reported for the CELF-4.

Overall, construct validity, including the reference standard, sensitivity and specificity, and likelihood ratios of the CELF-4 was determined to be insufficient due to potential spectrum bias in the standardization sample and lack of validity of the reference standard as well as unacceptable sensitivity and specificity measures.

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 validity. For example, consider a standard scale in the produce section of a grocery store. Say a consumer 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 consumers 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 (7 to 35 days; mean = 16 days) to 320 individuals from the standardization sample. Children aged 6-21 were used for this study, which does not include the entire age range of the CELF-4. Salvia, Ysseldyke, and Bolt, 2010, as cited in Betz, Eickhoff, and Sullivan, 2013, recommend that *minimum* standard for test reliability be .9 when using the test to make educational placement decisions, including SLP services. According to the Examiners Manual, across ages and subtests, reliability coefficients ranged from .71 to .86. Test-retest reliability is insufficient.

Inter-examiner Reliability– Inter-examiner reliability is used to measure the influence of 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 30 trained raters. Each subtest 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 .88 to .99 across seven selected subtests. All but one subtest (*Word Definitions*) met the standards in the field for reliability (Salvia, Ysseldyke, and Bolt, 2010, as cited in Betz, Eickhoff, and Sullivan, 2013). However, it is important to note that "a high degree of reliability alone does not ensure validity" (McCauley & Swisher, 1984, p. 35).

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 (5-21) and subtests, coefficients ranged from .71 to .92; 4 out of 12 subtests did not meet the standard for reliability. Inter-item consistency was also calculated for 405 students from four clinical groups: LD, intellectual disability, autism, and hearing impaired. Across subtests, the split-half coefficient ranged from .85-.97, indicating that the CELF-4 is equally reliable for measuring language skills of clinical groups and TD children.

Overall, the reliability, including the test-retest, and inter-examiner reliability, is considered insufficient. Specifically, across ages for the range of the test, none of the subtests received sufficient test-retest reliability as the coefficients were all below .9. Test-retest reliability was not calculated using the entire age range of the CELF-4, thus making it an insufficient measure. Further, one subtest did not meet the standard for inter-examiner reliability, and 4 out of 12 subtests did not meet the standard in the field for inter-item consistency to be considered sufficient.

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, intended 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 the age group of 15 years and 4 months of age on the CELF-4, the SEM at the 90% confidence level is +/- 6 for Core Language Score (CLS). If the child were to achieve an 80 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 74 and 86. Thus, all the clinician can determine from administration of the CELF-4 is that this child's true language ability (according to the CELF-4) ranges from moderately impaired to within normal limits. Although it varies, many schools consider children to have a LD if they fall below one SD from the mean (Semel, Wiig, & Secord, 2003), which would be less than 85 on the CELF-4. Without considering the CI, this child would be labeled LD inappropriately and given special education services unnecessarily. This has serious long term consequences on the child's development and achievement. The wide range of the CI makes the scores from the CELF-4 insufficient as children may be misdiagnosed.

7. BIAS:

According to Crowley (2010), IDEA 2004 regulations stress that assessment instruments must not only be "valid and reliable" but also free of "discriminat[ion]on a racial or cultural basis." In addition to being an invalid measure of language ability, the CELF-4 contains many inherent biases against culturally and linguistically diverse children.

Linguistic Bias

English as a Second Language

According to Paradis (2005), 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 (2004), clinicians working with diverse and bilingual backgrounds must be familiar with "how linguistic features and learning characteristics of language differences and second language acquisition are different from those associated with a true learning disability, emotional disturbance, central auditory processing deficit, elective mutism, or attention deficit disorder."

Take this example from the CELF-4 which demonstrates how ESL students may be falsely identified as having a language disorder on the *Formulated Sentences* subtest. In this subtest, students are shown a picture and given a target word. Then they are asked to use the word in a sentence. The first item in this subtest depicts children playing a video game with their father. A child learning English as a Second Language may produce "The children playing the video game," omitting the verb "are" in this present progressive construction, which is a common characteristic of children whose first language is Spanish (Shames, Wiig, & Secord, 1998). According to the Examiner's Manual, this response would be given a score of one out of two, and thus the child is not given full credit for providing a response that is simply consistent with second language acquisition.

Dialectal Variations

A child's performance on the CELF-4 may also be affected by the dialect of English that is spoken in their homes and communities. It is important to consider the dialect issues of the test being administered in Standard American English (SAE) with speakers of other dialects. 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 more difficult for a speaker of SAE to repeat this sentence as compared to a similar sentence in SAE.

Speakers of dialects other than SAE (e.g. African American English (AAE), Patois) face a similar challenge when asked to complete tasks such as the *Recalling Sentences* subtest of the CELF-4. 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 be noted that in the Examiner's Manual, accommodations for speakers of dialects other than SAE are suggested. For example, on page 312, alternate responses to the *Word Structure* items are given for speakers of AAE. Although the alternate responses are intended to add scoring sensitivity to speakers of dialects other than SAE, AAE is the *only* other dialect that is provided in the manual. Other dialects are not mentioned in the manual, and thus speakers of these dialects may be at a disadvantage. It should also be noted that speakers of dialects fall along a continuum; they may use features of both their native dialect *and* SAE. Examiners are not sensitive to dialectal issues may expect a child to use *only* one dialect, which is not usually the case. Further, in the scoring directions for the *Formulated Sentences* subtest, notes are provided on relevant items to give full credit for responses that are dialectal. For example, for item 2, a child would be given full credit for the response "Her forgot her mittens," which is a typical response for a speaker of AAE. When administering the CELF-4, it is important to note the alternate responses suggested in the manual when scoring tests of speakers of dialects other than SAE. However, despite recognition of dialect differences acceptable responses, the creators of the CELF-4 fail to recognize the bias inherent in administering a test in SAE to a speaker who may not be familiar or proficient in SAE.

Socioeconomic Status Bias

Hart & Risley (1995) found that a child's vocabulary correlates with his/her family's socio-economic status (SES); parents with low SES used fewer words per hour when speaking to their children than parents with professional skills and higher SES. Children from families of higher SES tend to have larger vocabularies and score better on standardized tests since many items are actually vocabulary based. A child from a lower SES background may be falsely identified as having a language disorder on standardized language tests due to a smaller vocabulary than his higher SES peers. The CELF-4 contains many items that are biased against children from low SES backgrounds because they require knowledge of lower frequency vocabulary items. For example, on the *Expressive Vocabulary* subtest, a child from a lower SES may not have exposure to some of the lower frequency vocabulary words such as trophy, skeleton, telescope, and binoculars. Also, the *Formulating Sentences* and *Word Classes 1 and 2* subtests require prior knowledge of the stimulus word to provide an appropriate response. As a result of vocabulary items on the CELF-4, children from low SES backgrounds will likely have reduced scores when compared to higher SES peers.

Prior Knowledge/Experience

A child's performance on the CELF-4 may also be affected by their prior knowledge and experiences. For example, a child from the middle of the country may not have prior experience with the word *island* in the *Expressive Vocabulary* subtest; a child who has never attended school and interacted with school supplies would not be familiar with the objects in item 21 from the *Word Classes I* subtest (eraser, glue, chalk, tape).

It is also important to consider that the format of the test may affect a child's performance if they do not have prior experiences with the specific type of testing. According to Peña, &

Quinn (1997), children from culturally and linguistically diverse backgrounds do not perform as well on assessments that contain tasks such as labeling and known information questions , as they are not exposed to these tasks in their culture. The CELF-4 contains various testing formats, many of which are dependent upon prior knowledge and experience. The *Expressive Vocabulary* subtest is a task based entirely on labeling. A child who has not been exposed to this type of testing may label a shoe as “walking” as they have been exposed to function-type description tasks rather than labeling the object itself. Further, the *Understanding Spoken Paragraphs* subtest requires the child to respond to a “known information question.” In this subtest, the student is required to listen to a short passage read aloud by the clinician and then respond to questions related to the story. A child who is not accustomed to an adult asking them questions to which they already know the answer may fail to respond appropriately.

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

Cultural Bias

According to Peña & Quinn (1997), tasks on language assessments often do not take into account variations in socialization practices. For example, the child’s response to the type of questions that are asked (e.g. known 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. 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.

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.

The CELF-4 suggests standard administration to include all 17 subtests. Such administration would require the child to be exceptionally motivated and interested. Further, fatigue may affect performance on later items in the test's administration. Should all 17 subtests be administered, subtests at the end are likely to be negatively impacted by the child's fatigue.

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. 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.

8. **SPECIAL ALERTS/COMMENTS**

The CELF-4 was designed to assess the presence of a language disorder or delay in children aged 5-12;11 using a comprehensive and flexible assessment approach. Subtests were designed in correspondence with educational mandates with regards to (a) eligibility for services, (b) identification of strengths and weaknesses, and (c) performance within tasks related to the standard educational curriculum. Despite the CELF-4'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, an insufficient reference standard, and insufficient accuracy in determining the presence of a language disorder. Despite these discrepancies in validity, the CELF-4 touts "fair" to "good" sensitivity and specificity data when compared to standards in the field; however, these numbers are misleading as the populations chosen for the sensitivity (language disordered) populations were chosen through an invalid measure (e.g. previous versions of the CELF lacking in acceptable diagnostic accuracy). Therefore, even if the CELF-4 were a valid, reliable and unbiased assessment, it lacks sufficient discriminant accuracy in order to determine the presence or absence of a language disorder.*

In addition, as this test is largely a test of vocabulary based in prior knowledge and experience, for second language learners including English-dominant 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.

Due to cultural and linguistic biases (for example, exposure to books, repetition of unfamiliar syntactic structures if not proficient in Standard American English, cultural labeling practices, communication with strangers, responses to known questions, etc.) and assumptions about past knowledge and experiences, this test should only be used to probe for information and not to identify a disorder or disability. 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. They recommend modifications that may be used in test administration to account for cultural and linguistic differences including increasing response time, increasing the number of trial items, supporting test results with language sampling, and observing the student in the classroom (Semel, Wiig, & Secord, 2003). A complete list of modifications is available on page 11 of the Examiner's Manual. If any such modifications are used, the authors caution that normative test scores cannot be used, and a descriptive approach should be utilized in reporting the student's responses. Therefore, scores should not be calculated nor should they be used for 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.

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

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.

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.

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.

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

Shames, G. H., Wiig, E. H., & Secord, W. A. (1998). *Human Communication Disorders: An Introduction (5th ed.).* Boston, MA: Allyn and Bacon.

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.