Test Review:
Preschool Language Scales- Fifth Edition (PLS-5)

Version: 5th edition
Copyright date: 2011
Grade or Age Range: Birth-7;11
Authors: Irla Lee Zimmerman, Ph.D., Violette G. Steiner, B.S., & Roberta Evatt Pond, M.A.
Publisher: Pearson

Section	*Page Number*
1. Purpose	pg. 2
2. Description	pg. 2
3. Standardization Sample	pg. 3
4. Validity	pg. 3
a. Content	pg. 4
b. Construct	pg. 4
1. Reference Standard	pg. 5
2. Sensitivity and Specificity	pg. 6
3. Likelihood Ratio	pg. 7
c. Concurrent	pg. 8
5. Reliability	pg. 9
a. Test-Retest Reliability	pg. 9
b. Inter-examiner Reliability	pg .9
c. Inter-item Consistency	pg. 10
6. Standard Error of Measurement	pg. 11
7. Bias	pg. 11
a. Linguistic Bias	pg. 11
1. English as a Second Language	pg. 11
2. Dialectal Variations	pg. 12
b. Socioeconomic Status Bias	pg. 13
c. Prior Knowledge/Experience	pg. 13
d. Cultural Bias	pg. 15
e. Attention and Memory	pg.15
f. Motor/Sensory Impairments	pg. 17
8. Special Alerts/Comments	pg.17
9. References	pg. 18

---

1. PURPOSE

The PLS-5 is designed for use with children aged birth through 7;11 to assess language development and identify children who have a language delay or disorder. The test aims to identify receptive and expressive language skills in the areas of attention, gesture, play, vocal development, social communication, vocabulary, concepts, language structure, integrative language, and emergent literacy (Examiner's Manual, pg. 3). The PLS-5 aids the clinician in determining strengths and weaknesses in these areas in order to determine the presence and type of language disorder (e.g. receptive, expressive, and mixed), eligibility for services and to design interventions based on norm-referenced and criterion referenced scores. Although it is not intended to determine if a child is gifted, it may provide appropriate supplemental information regarding their language development.

2. DESCRIPTION

The PLS-5 consists of two standardized scales: Auditory Comprehension (AC), to "evaluate the scope of a child's comprehension of language," and Expressive Communication (EC), to "determine how well a child communicates with others"(Examiner's Manual, pg. 4). Administration time varies based on the child's age and can range between 25-35 minutes for children aged birth through 11 months to 45-60 minutes for children over one year. Specific AC tasks assessed include comprehension of basic vocabulary, concepts, morphology, syntax, comparisons and inferences, and emergent literacy. Specific EC skills include naming, describing, expressing quantity, using specific prepositions, grammatical markers, sentence structures, and emergent literacy skills. Three optional supplemental measures are also included (Language Sample Checklist, Articulation Screener, and Home Communication Questionnaire). Scores are provided at three month intervals from birth through 11 months, and at 6 months intervals from 1 year through 7;11. The PLS-5 yields norm-referenced scores including standard scores, percentile ranks and age equivalents for the AC and EC scales as well as for Total Language (TL). However, the manual warns against the use of age equivalent scores as this type of score does not provide the sufficient information to determine the presence of a language disorder, can be easily misinterpreted and have a number of psychometric limitations (Examiner's Manual pg. 17) The test recommends that examiners calculate norm-referenced scores to identify speech and language disorders. However the test does comment that evaluation of a child can also include portfolio assessment, dynamic assessment and parent/caregiver interview (Examiner's Manual, pg. 9). Caregiver's responses to the Home Communication Questionnaire may support items on the AC and EC scales from birth through 2;11. According to the test manual, the PLS-5 may only be administered by trained professionals including speech-language pathologists, early childhood specialists, psychologists and other professionals who have training and experience in diagnostic assessment of children of this age.

3. STANDARIZATION SAMPLE

The standardization sample for the PLS-5 included 1400 children aged birth through 7;11. The standardization sample was matched to the 2008 United States census and was stratified by demographic factors including age, sex, geographic region, race/ethnicity, and primary-care giver's highest education level. Inclusion into the standardization sample required completion of the test without modifications. English was required to be the primary language for all subjects for both comprehension and expression. For preverbal children, English was the primary language of the caregivers in the home. Approximately 3% of the sample population was from homes that spoke a language other than English. No note was made in the Examiner's Manual to match the standardization sample to U.S. census data regarding children who spoke languages other than English in the home. The standardization sample consisted mainly of children who spoke SAE (78.9%). The sample also included 4.2% of children who spoke African American English (AAE), 5.8% who spoke Spanish influenced English, 4.4% who spoke Southern English, and less than 3% who spoke other dialects. Scoring rules were adapted for children who spoke AAE, Spanish-influenced English, Chinese-influenced English, Appalachian English and Southern English so that children would not be penalized on test items that assess dialect specific linguistic skills. However these modified rules only accounted for a portion of the other dialects in the sample. No information is included in the manual explaining how participants were selected. The manual does not discuss whether participants with disabilities were included in the standardization sample. This is relevant because inclusion of participants with disabilities in the standardization sample lowers the mean score of the test and negatively impacts the test's ability to distinguish between typically developing children and children with disorders (Pena, Spaulding & Plant, 2006).

4. VALIDITY

Content - Content Validity refers to how representative the test items are of the content that is being assessed (Paul, 2007). Content validity was analyzed using literature reviews, clinician feedback, expert review and response processes. New items on the PLS-5 were refined from the PLS-4 to reflect current research on language development and were determined via literature review and clinician feedback. Children's response processes and clinician feedback during the pilot and tryout phases of the PLS-5 development was used to ensure the appropriateness and breadth of test items. Tryout testing to evaluate the appropriateness and breadth of the PLS-5 took place between February and July 2009. Two samples were collected: a nonclinical sample of 455 children aged 0-7;11 who had not been previously diagnosed with a language disorder and a clinical sample of 169 children aged 2-7;11 diagnosed with a receptive or expressive language disorder based on a score of 1.5 SD below the mean on an unspecified standardized language test. Since we are unable to

evaluate the accuracy and validity of the language tests used to classify the clinical sample, the standardization process merely determined if the PLS-5 scores of the children matched their scores on other unspecified language tests. It did not determine the presence of a disability and the clinical sample's true diagnostic status is unknown. As well, according to Spaulding, Plante and Farinella (2006), the practice of using an arbitrary cut-off score to determine disability is unsupported by the evidence and increases the chances of misdiagnosis. Currently, no commercially available test is considered acceptably accurate in identifying a disorder based on a score alone and research demonstrates that standardized language tests do not consistently diagnose children correctly (Dollaghan and Horner, 2011).

Items were revised or deleted if they did not sufficiently differentiate between the clinical and nonclinical samples or if the items were considered unfair or too difficult to score. PLS-5 content was assessed for bias during pilot and tryout testing via a review of a panel of experts with experience in assessment issues related to cultural and linguistic diversity. The panel included speech-language pathologists and one psychologist who are professors at various universities in the United States. Specific information regarding the background and training of the "panel of experts" was not provided. As a result, the expert review panel may have been limited in its ability to accurately assess the test content for bias. According to ASHA (2004), clinicians working with culturally and linguistically diverse clients must demonstrate native or near-native proficiency in the language(s) being used as well as knowledge of dialect differences and their impact on speech and language. It is unknown if this panel of experts was highly proficient in the variety of dialects and complexity of linguistic differences for which they were evaluating content. Therefore we cannot be certain that test items are free from cultural and linguistic biases. Due to lack of information regarding method of selection of sample populations and diagnosis of the clinical population as well as the training and background of "the expert" panel, content validity of the PLS-5 cannot be considered sufficient.

Construct – Construct validity assesses how well the test measures what it purports to measure (Paul, 2007). It was measured by comparing the performance of special groups of children with language disorders or delays to typically developing children. The TD children were defined as children who had not been previously diagnosed as having a language disorder and who were not receiving speech and language services at the time. The children with a language disorder or delay were defined based on a score of 1.5 SD below the mean on an unspecified language test. The diagnosis of each group of children was compared with their status to determine the diagnostic accuracy of the PLS-5. Once again, the lack of information regarding what measure was used to determine diagnostic status immediately calls into question the construct validity of the PLS-5. Also, as mentioned previously, the use of an arbitrary cut score has been demonstrated not to be an effective or accurate way to determine disability (Spaulding, Plante and Farinella, 2006). Clinical samples identified

through the use of arbitrary cut off scores should not provide evidence for construct validity and diagnostic accuracy.

Reference Standard

In considering the diagnostic accuracy of an index measure such as the PLS-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 for the sensitivity group was a score below 1.5 SD on an unspecified standardized test of language skills. The reference standard was applied to two groups of children to determine the sensitivity measure. One group was classified as language disordered (LD) and consisted of 229 children at least three years of age who scored at least 1.5 SD below the mean on an unspecified language test and were enrolled in a language therapy program at the time of test administration. The second group was classified with developmental language delays (DLD) and consisted of 23 children between one year and 3;11 who scored at least 1.5 SD below the mean on an unspecified standardized test of language skills and were enrolled in a language stimulation program. The test manual does not specify the language tests used to classify the groups of affected children. As well, the test manual does not explain why, other than age range, a distinction was made between the two clinical groups since they only differentiating factor is age. Therefore, the validity of these tests is unknown and we are unable to determine the accuracy of these tests in identifying children with language disorders or delays.

It should be noted that children included in the DLD or LD groups were classified with moderate to severe language delays. Children with mild language delays were not included in the study (Examiner's Manual pg. 93). In fact, to better distinguish between children with a developmental language delay and typically developing children the distinguishing score was shifted from 1 SD (cut score of 85) to 1.5 SD (cut score of 77). The authors noted that the initial inclusion criteria were amended because at scores of 1 SD below the mean it was difficult to distinguish children with mild language delays from those children that were typically developing (pg. 93). This inflates the diagnostic accuracy of the test because it does not reflect the test's ability to distinguish between TD and children with mild DLD. Therefore, the diagnostic accuracy reported by the PLS-5 demonstrates a 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).

Many issues result in insufficient construct validity for the PLS-5. The reference standard was not identified and therefore could not be evaluated. Additionally, the reference standard used was not applied to the children classified as typically developing therefore we cannot be sure they are free from the disorder (Dollaghan, 2007). This affects the base rate, sensitivity and specificity measures and likelihood ratios. Construct validity is reduced because test designers excluded children with mild language disorders in order to inflate the diagnostic accuracy reported in the test manual.

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 an eight-year-old boy previously diagnosed with a language disorder, will achieve a score indicative of having a language disorder on this assessment. For the group of children identified as LD (ages 3;11-7;11), the PLS-5 reports the sensitivity to be .83 at a cut score of 1 SD or more 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 of this measure would be considered "fair." It is important to consider the implication of this measure; a sensitivity of .83 means that 17/100 children with a language disability will be identified as typically developing and will not receive appropriate services. For the group of children identified as DLD (ages 0-3;11), the PLS-5 reports the sensitivity to be .91 at a cut score of 1 SD below the mean. This measure would be considered "good" according to the standards in the field (Plante & Vance, 1994). However, the reported measures are invalid due to spectrum bias noted in the typically developing and language delayed/disordered group. Additionally, because the reference standard was previously determined to be invalid or unknown, it is also unknown whether the sensitivity measures actually reflect the test's diagnostic accuracy.

Specificity measures the proportion of typically developing students 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. In the clinical study with the group of TD children, the PLS-5 reports specificity measures to be 0.8 at a cut score of 1 SD below the mean, which would be considered a "fair" measure according to the standards in the field (Plante & Vance, 1994). It is important to consider the implications. A specificity of .8 means that 20/100 typically developing children will be identified as having a language disorder and may be inaccurately referred for support services. In the clinical study with the group of DLD children, the PLS-5 reports specificity measures to be .78, an unacceptable measure according the standards in the field (Plante & Vance, 1994).

Additionally, because the reference standard was previously determined to be invalid or unknown, the specificity measures do not reflect the test's diagnostic accuracy.

The same reference standard was not applied to both the specificity and sensitivity groups. This decreases the validity of the test due to spectrum bias which occurs when the sample population does not represent the full spectrum of the clinical population (Dollaghan & Horner, 2011). Due to lack of information about the reference standard, the diagnostic status of the specificity group is unknown so we cannot be sure about the accuracy of the specificity measure. Sensitivity and specificity were also determined to be unacceptable, despite reported measures considered to be "fair" or "good" (.80 or above), due to lack of information regarding the reference standard as well as spectrum bias caused by different reference standards being used for different clinical populations.

Likelihood Ratio

According to Dollaghan (2007), likelihood ratios are used to examine how accurate an assessment is at distinguishing individuals who have a disorder from those who do not. A positive likelihood ratio (LR+) represents the likelihood that an individual who is given a positive (disordered) score on an assessment actually has a disorder. The higher the LR+ (e.g. >10), the greater confidence the test user can have that the person who obtained the score has the target disorder. Similarly, a negative likelihood ratio (LR-) represents the likelihood that an individual who is given a negative (non-disordered) score actually does not have a disorder. The lower the LR- (e.g. < .10), the greater confidence the test user can have that the person who obtained a score within normal range is, in fact, unaffected.

Likelihood ratios for the reference standard are not reported as the reference standard was not applied to the typically developing group. While the reference standard was applied to the language disordered group and the language delayed group, the PLS-5 does not report how many children in each group scored below a score of 77 (the cut-off score). Thus, the sensitivity value does not truly reflect the test's diagnostic accuracy and consequently likelihood ratios cannot be calculated.

Overall, construct validity, including the reference standard, sensitivity and specificity, and likelihood ratios of the PLS-5 was determined to be insufficient. An unspecified reference standard invalidates the diagnostic accuracy of the test because we cannot be sure children identified as having a disability by the reference standards were accurately diagnosed. As well, because no reference standard was applied to the non-clinical group for the specificity measure we cannot be sure these children are free from a disorder. Inflated diagnostic accuracy reported in the test manual contributes to concern regarding the validity of the PLS-5 in detecting the presence of language disorder. In addition, the authors changed the diagnostic criteria for distinguishing between LD and TD individuals to below 1.5 SD when they realized that scores below 1 SD did not distinguish between the two groups. This produces spectrum bias because it excludes those individuals who have a mild disorder. As a result, the population used to

determine the validity of the PLS-5 does not represent the clinical population encountered by speech language pathologists, which is likely to include children with mild language delay or disorder. This makes the PLS-5 inappropriate for real-world applications and intentionally leads an evaluator who has not carefully read the manual to believe the PLS-5 has a level of accuracy which it does not possess when applied to real clinical populations. The test manual does not state that it is only intended to identify children as moderately or severely delayed. In addition, in determining concurrent validity, the reference standards used (below 1.5 SD on the PLS-4 or CELF-P2) were themselves invalid measures for determining the presence of a language delay or disorder and these comparison populations did not cover the entire age range for which the PLS-5 was designed. Therefore, the diagnostic accuracy of the PLS-5 is insufficient and the PLS-5 cannot be considered a valid diagnostic tool.

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 another test 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 performance of a clinical sample on the PLS-5 to two other child language assessments: the CELF-P2 and the PLS-4. The study conducted to compare the PLS-5 to the PLS-4 consisted of a sample of 134 children aged 0-6;11 as this is the age range for the PLS-4. Correlation coefficients for the study were .80 for the AC and EC scales and .85 for TL. The study conducted to compare the PLS-5 with the CELF-P2 consisted of a sample of 97 children aged 3-6;11 as this is the age range for the CELF-P2. Correlation coefficients for the study ranged between .70-.82. According to Salvia and Ysseldyke (as cited in McCauley and Swisher, 1984), a correlation coefficient of .90 or better is needed to provide sufficient evidence. It is important to note that the entire age range for which the PLS-5 is intended was not standardized. According to the test manual, the PLS-5 is intended to diagnose language delay/disorder in children from birth to 7:11, however concurrent validity was not determined for ages 6;11- 7;11. 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 PLS-4 has a specificity measure of .90 but a sensitivity measure below .50 which is unacceptable according to the standards in the field and should not be used to determine the concurrent validity of the PLS-5. The CELF-P2 had a sensitivity of .85 and specificity of .82 on the core language score indicating fair diagnostic accuracy. However, since it only compares children between 3-6;11 it does not account for children in the range for which the PLS-5 is intended (0-7;11) it should not be used as a comparison measure. Due to comparison measures which do not meet acceptable levels of validity and/or measures which do not cover the entire age range for which the PLS-5 is intended, concurrent validity was found to be 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 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 good *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 good *inter-examiner reliability.* 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.

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 twice to 195 children from the normative sample ranging in age from birth to 7;11. The administration was conducted by the same examiner, and the testing interval ranged from 3-28 days. Correlation coefficients were calculated for Auditory Comprehension, Expressive Communication, and Total Language for three age brackets: 0;0-2;11, 3;0-4;11, and 5;0-7;11, yielding nine correlation coefficients. The reliability coefficients ranged from .86 to .95. 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. As well, the small sample size of children in each age band limits the reliability measure. Thus, the test-retest reliability for the PLS-5 is considered insufficient due to a small sample sizes and because three out of nine correlation coefficients were less than the accepted minimum standard.

Inter-examiner Reliability– Inter-examiner reliability is used to measure the influence of different test scores or different test administrators on test results (McCauley & Swisher, 1984). It should be noted that the inter-examiner reliability for index measures is often calculated using specially trained examiners. When used in the field, however, the average clinician will likely not have specific training in test administration for that specific test and thus the inter-examiner reliability may be lower in reality. Inter-examiner reliability was assessed in a study where two examiners assessed 54 children in two age brackets: birth - 3;11 and 4;0 - 7;11. Inter-examiner reliability coefficients ranged from .96 - .99 across

subtests for both age groups, indicating acceptable inter-examiner reliability (Salvia, Ysseldyke, & Bolt, 2010, as cited in Betz, Eickhoff, & Sullivan, 2013).

Test developers additionally conducted a second study, referenced as the interscorer study, to evaluate the consistency of scoring rules for test items that are judged subjectively or where there is room for interpretation. For example, item 38 on the EC scale requires the child to answer questions logically. To be considered correct a specific answer is not required. Rather, the examiner judges the item to be correct based on their interpretation of the child's answer. To examine inter-scorer reliability, scores were compared between trained scorers and the examiner to determine if scoring rules were clear and objective. Trained scorers consisted of five individuals who had been trained in applying scoring rules to the subjective items. Two hundred test protocols were randomly selected for use in this study. Interscorer agreement ranged from 91.9 to 100%, indicating acceptable reliability (Salvia, Ysseldyke, & Bolt, 2010, as cited in Betz, Eickhoff, & Sullivan, 2013). This implies that most clinicians will arrive at the same score decision for items that have subjective scoring and that interscorer reliability meets acceptable standards.

Inter-item Consistency – Inter-item consistency assesses whether "parts of the test are measuring something similar to what is measured by the whole" (Paul, 2007). Inter-item consistency was calculated using a split half coefficient for three populations: the normative sample, children with language disorders and children with language delays. Correlation coefficients ranged between .91 and .98 for all three groups. Please see the chart below for correlation coefficients.

Population	AC Coefficient	EC Coefficient	TL Coefficient
Normative Sample	.91	.93	.95
Children with Language Disorders	.97	.97	.98
Children with Language Delays	.96	.93	.97

Based on these numbers, inter-item consistency is considered acceptable (Salvia, Ysseldyke, & Bolt, 2010, as cited in Betz, Eickhoff, & Sullivan, 2013).

Overall, most types of reliability for the PLS-5, including inter-item consistency and inter-examiner reliability, meet standards established as acceptable for the field. Test-retest reliability was compromised, however, due to unacceptable coefficients and issues regarding sample size and short latency time between test-retest administrations that could lead to a practice effect. Regardless, "a high degree of reliability alone does not ensure validity" (McCauley & Swisher, 1985, pg. 35). As noted in previous paragraphs, the PLS-5 was not found to be a valid instrument for identifying the presence of language disorder or delay.

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 standard score, which suggests the range in which their "true" score falls. Children's performance on standardized assessments may vary based on their mood, health, and motivation. For example, a child may be tested one day and receive a standard score of 90. Say he was tested a second time and he was promised a reward for performing well; he may receive a score of 96. If he were to be tested a third time, he may not be feeling well on that day, and thus receive a score of 84. As children are not able to be assessed multiple times to acquire their "true" score, the SEM and CIs are calculated to account for variability that is inherent in individuals. Current assessment guidelines in New York City require that scores be presented within 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 confidence interval. Although a larger range of scores is yielded with a higher confidence interval, the clinician can be more *confident* that the child's 'true' score falls within that range. A lower level of confidence will produce a smaller range of scores but the clinician will be less confident that the child's true score falls within that range. The wide range of scores necessary to achieve a high level of confidence, often covering two or more standard deviations, demonstrates what little information is gained by administration of a standardized test.

The PLS-5 provides CIs at the 90% and 95% confidence levels for AC, EC, and TL. For example, consider a child aged 2;0-2;5 who received a raw score of 22 on the EC subtest, that raw score converts to a standard score of 77. As this score falls 1.5 SD from the mean, this child would likely be classified as having a mild expressive language disorder. However, according to the manual, at a 90% confidence level, the child's true score falls between 72 and 87. The lower bound of this interval would suggest a moderate to severe expressive language impairment while the upper bound would classify the child as typically developing. We cannot determine eligibility for special education services and provide a disability label based on a measure with this much variability (even if the test were a valid measure).

7. **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 with limited English proficiency (LEP) 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 AC subtest, children are presented with a grammatically incorrect sentence such as, "Her can eat cookies" (pg. 52). Children are required to make grammaticality judgments and provide the correct response. A child with LEP, who has difficulty with pronouns, may respond with "Her ate cookies" which would be considered incorrect. The answer however may reflect the child's lack of exposure to English pronouns rather than a disorder since an acceptable response to this sentence could be "she ate cookies". As well, many items in the EC subtest are judged correct based on the grammaticality of the child's response. For example, a child is required to correctly describe pictures using present progressive tense or formulate grammatically correct questions in response to picture stimuli. All of these items could be challenging for a child with LEP and they could be misdiagnosed as having a language disability. Despite research demonstrating the similarity in expressive language of LEP children and LD children, the manual makes no mention of this fact to alert clinicians to the potential inappropriateness of the test.

Dialectal Variations

A child's performance on the PLS-5 may also be affected by the dialect of English that is spoken in his or her home and community. It is important to consider the dialect issues caused by 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, it would be difficult for a speaker of SAE to repeat this sentence.

Speakers of dialects other than SAE (e.g. African American English (AAE), Patois) face a similar challenge when asked to complete tasks such as sentence repetition e.g. "When he came home from school, Joey ate a snack", EC #57), story retelling (AC #53-57, EC #58-60) and sentence formulating (EC #61). 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. As a result, the child's performance will necessarily be affected and is not comparable to a normative sample consisting almost entirely of SAE speakers. The PLS-5 takes into consideration dialectal variations in scoring but does not account for the variety of dialects spoken in the United States. As well, if the examiner does not speak the same dialect it may place an extra burden on the child or they may be uncomfortable using their dialect and attempt to switch to SAE despite being less proficient. The test designers fail to recognize this inherent bias and do not suggest accommodations that examiners can make or even acknowledge this bias in the manual.

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 (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. A child from a lower SES background may be falsely identified as having a language disorder if they are unable to correctly label items or answer questions that rely on vocabulary exposure such as the understanding analogies items in the AC subtest. Children from low SES often perform more poorly than age-matched peers from middle SES on standardized language tests. In contrast, there is often no difference in the abilities of these two groups to learn novel words, demonstrating the bias of standardized tests against children from low SES as these tests are highly correlated to SES (Horton-Ikard & Weismer, 2007). According to Peña and Quinn (1997), children from a low SES have less exposure to labeling and may be at a disadvantage in tasks that require labeling such as picture identification on the AC scale [e.g. television (#31), rollerblades (# 46)] or picture naming in the EC scale [e.g. balloon (#26)]. A child from a lower SES may not have had exposure to certain items such as a rollerblades, which would impair their ability to correctly answer these questions. According to Hart and Risley (1995), students from low SES on average have significantly less exposure to vocabulary and less familiarity with books. They would be especially challenged on the story retelling and would be less able to demonstrate emergent literacy through book handling and print awareness on the AC scale (#63).

Prior Knowledge/Experience

A child's performance on the PLS-5 may also be affected by their prior knowledge and experiences. For example, a child from a city may not know what a frog (AC #51) or motor boats (AC #58) are because they have not been exposed to these things previously.

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 PLS-5 contains a number of tasks that are dependent on prior knowledge and experience. Many items on the EC scale are based entirely on labeling. A child who has not been exposed to this type of testing may not perform well on these tasks. For example, EC item 30 requires the child to label items as the examiner points to various pictures [e.g. cookie, scissors, banana]). A child may label *scissors* as "used for cutting" as they have been exposed to function-type description tasks in their culture rather than labeling the object itself. According to Pena and Quinn (1997), children from different cultures tend to use functional descriptions to label objects based on culturally dependent maternal teaching strategies and different cultural expectations. As a result, he or she may be falsely identified as having a language disorder due to poor performance on similarly culturally dependent tasks such as this one.

Further, many tasks on the PLS-5 require the child to respond to a "known question," where the clinician obviously knows the answer. The child and the clinician attend to a single image and the child is required to provide information. For example, on the "understands negatives in sentences" portion of the AC subtest, the clinician provides a probe and the child responds by pointing, (e.g. "Look at all the babies. Show me the baby who is not crying"). If the child has not been exposed to this type of question, they may answer incorrectly because they are unfamiliar with what is expected of them and their performance will not reflect their true skills.

Further, a child's performance on the test may have been affected by their prior exposure to books and toys. According to Peña and Quinn (1997), some infants are not exposed to books, print, take-apart toys, or puzzles. The PLS-5 requires the child to interact appropriately with different toys including blocks, a teddy bear, and various kitchen items. If the child has no prior experience with these toys or this type of play, their performance may not reflect their true skills. The PLS-5 demands the child attends to the test book for the length of the assessment, something he or she may be unaccustomed to doing. He or she must also realize that pictures and symbols have meaning and have familiarity with how to attend to them. This is referred to as print awareness, which is a learned skill. Lack of prior exposure to books and print materials may result in a lack of familiarity with letters and sounds. This can appear to be a delay in pre-literacy skills including phonological awareness and letter knowledge when compared to a normative sample largely made up of children with adequate and mainstream exposure to print and books. This may negatively affect the child's performance on the AC scale when they are required to identify word initial sounds (#51) or rhyme words on the EC scale (#53).

Cultural Bias

According to Peña & Quinn (1997), tasks of 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. For example, 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 to 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. Also, a child's level of comfort making eye contact with the examiner or speaking to an adult may affect their performance on items in the EC subtest that require the child to elaborate on or retell a story or formulate sentences.

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 the 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.

PLS-5 administration time varies depending on the child's age, ability, attention span and willingness to participate during the test. Administration time for young (birth – 11mo) children ranges between 25-35 minutes and increases up to an hour for children 3 years to 7 years 11 months. This can be taxing for children, particularly if they are struggling with the test or are uncomfortable with the testing situation. The test manual does permit break time for children to rest, have a snack or go to the bathroom in order to optimize their attention and focus. However, even with these concessions, many children (even those from mainstream, middle class backgrounds) will still be highly resistant to finishing the test or performing to their best ability due to lack of interest and/or motivation in the testing materials.

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. Further, as the sample population did not include children with attention and/or memory deficits, 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 they are not using proper accommodations, they may not be able to fully see the test stimuli, and thus their performance may not reflect their true abilities. A child with motor deficits, such as a child with typical language development but living with cerebral palsy (CP), may find it much more frustrating and tiring to be pointing to/attending to pictures for an extended period of time than a typically developing non-disabled child. The child with CP may not perform at his highest capacity due to his motor impairments and would produce a lower score than he or she is actually capable of achieving. Further, as the sample population did not include children with motor/sensory impairments, results of this assessment are invalid for children with motor and sensory impairments.

8. **SPECIAL ALERTS/COMMENTS**

The PLS-5 was designed to assess receptive and expressive language abilities in children aged 0-7;11 in order to determine the presence of a language delay or disorder. The test consists of an auditory comprehension scale and expressive communication scale to evaluate specific areas of strength and weakness. However, results obtained from administration of the PLS-5 are not valid due to an insufficient reference standard and insufficient discriminant accuracy to properly identify children with a language delay or disorder. Furthermore, there are other issues of concern that call into question the validity of the PLS-5 as a diagnostic measure of speech and language ability. The PLS-5 manual states that it is a valid test with sensitivity values that appear to be fair or acceptable according to the standards in the field. However, it is critical to note that these numbers are misleading as the clinical populations chosen for the sensitivity measure were chosen through an unspecified measure. Specificity measures reported as fair were invalid since no reference standard was applied to the non-clinical sample for the specificity measure. Additionally, the clinical sample used to determine sensitivity and specificity was subject to spectrum bias. Only those individuals who were clearly language disordered/delayed were included in the sensitivity sample and only those who were not receiving speech and language services were included in the specificity sample. Mildly delayed children and children who were borderline typically developing were excluded from diagnostic accuracy measures reported. Diagnostic accuracy measures are irrelevant as they were not based on a population representative a real world clinical population.

In addition to unacceptable measures of validity, accuracy and reliability in determining the presence of language disorder or delay, the PLS-5 contains significant cultural and linguistic biases which preclude it from being appropriate for children from diverse backgrounds. This is noted in ***federal*** *legislation which requires testing materials to be "valid, reliable and free of significant bias"(IDEA, 2004). As many parts of this test are largely vocabulary based, it will likely falsely identify children as language delayed or disordered from non-mainstream cultural and linguistic backgrounds or who come from lower socioeconomic status.*

Labeling children as disabled and placing them in special education when they do not need to be there has many long lasting and detrimental consequences. These consequences include a limited and less rigorous curriculum (Harry & Klingner, 2006), lowered expectations which can lead to diminished academic and post-secondary opportunities (National Research Council. 20002; Harry & Klinker, 2006) and higher dropout rates (Hehir,2005) Due to cultural and linguistic biases such as vocabulary and labeling tasks as well as assumptions about prior knowledge and experiences (Hart & Risley, 1995; Peña and Quinn, 1997), this test should only be used to probe for information and not to identify a disorder or disability. Even for children from mainstream, SAE speaking backgrounds, the test has not demonstrated adequate validity and diagnostic accuracy. Therefore, scores should not be calculated to determine classification or referral to special education services. A speech and language evaluation should be based on clinical observations, consideration of the child's prior experiences and development history, as well as the parental report. Performance should be described in order to gain the most accurate conclusions about the nature and extent of language skills and to develop appropriate treatment recommendations (McCauley & Swisher, 1984).

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.

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.

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.

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

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

Zimmerman, I. L., Steiner, V, G., & Pond, E. (2011). *Preschool Language Scales- Fifth Edition (PLS-5).* San Antonio, TX: Pearson.