**Test Review:**

**Bayley Scales of Infant and Toddler Development: Third Edition**

 Version: Third Edition

Copyright date: 2006

Grade or Age Range: 1-42 months

Author: Nancy Bayley

Publisher: PsychCorp

|  |  |
| --- | --- |
| ***Section*** | ***Page Number*** |
| 1. Purpose | pg. 2 |
| 2. Description | pg. 2 |
| 3. Standardization Sample | pg. 3 |
| 4. Validity | pg. 4 |
|  a. Content | pg. 4 |
|  b. Construct | pg. 5 |
|  1. Reference Standard | pg. 6 |
|  2. Sensitivity and Specificity | pg. 6 |
|  3. Likelihood Ratio  | pg. 7 |
|  c. Concurrent | pg. 7 |
| 5. Reliability | pg. 8 |
|  a. Test-Retest Reliability | pg. 8 |
|  b. Inter-examiner Reliability | pg. 9 |
|  c. Inter-item Consistency | pg. 9  |
| 6. Standard Error of Measurement | pg. 10 |
| 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. 12 |
|  c. Prior Knowledge/Experience  | pg. 13 |
|  d. Cultural Bias | pg. 13 |
|  e. Attention and Memory  | pg. 14 |
|  f. Motor/Sensory Impairments | pg. 15 |
| 8. Special Alerts/Comments | pg. 14 |
| 9. References | pg. 17 |

1. **PURPOSE**

The Bayley Scale of Infant and Toddler Development (Bayley-III) is designed to assess the developmental functioning of infants and young children 1-42 months of age. The primary purpose of the Bayley-III is to identify suspected developmental delays in children through the use of norm-referenced scores and to provide information in order to plan appropriate interventions rooted in child development research. In addition, the Bayley-III can be used to monitor a child’s progress during intervention and helps to develop an understanding of a child’s strengths and weaknesses in relation to five developmental domains: cognitive, language, motor, social-emotional and adaptive behavior. The Bayley-III has a flexible administration format to accommodate variability in the child’s age and temperament. However, deviations for the standard procedures, such as phrasing or repeated presentation of a test item, invalidate the use of norms according to the test manual.

1. **DESCRIPTION**

The Bayley-III consists of five scales. Three scales, the Cognitive Scale, the Language Scale and the Motor Scale are administered by the clinician. Two scales, the Social-Emotional Scale and the Adaptive Behavior Scale from the Social-Emotional and Adaptive Behavior Questionnaire, are completed by the parent or primary caregiver. The Cognitive Scale assesses: play skills; information processing (attention to novelty, habituation, memory, and problem-solving); counting and number skills. The Language Scale contains receptive and expressive language subtests to assess communication skills including language and gestures. The Motor Scale is divided into the Fine Motor subtest and Gross Motor subtest. The Social-Emotional Scale assesses emotional and social functioning as well as sensory processing. It is based on The *Greenspan Social-Emotional Growth Chart: A Screening Questionnaire for Infants and Young Children* (The *Greenspan*; Greenspan, 2004). The Adaptive Behavior Scale assesses the attainment of practical skills necessary for a child to function independently and meet environmental demands. It is based on the *Adaptive Behavior Assessment System-Second Edition* (ABAS-II; Harrison & Oakland, 2003). The only modifications to The *Greenspan* and ABAS-II in the Bayley is the use of scaled scores in addition to the originally provided cut scores so that these measures may be more easily compared to the other Bayley-III subtest scores.

The Bayley-III provides norm-referenced scores. Scaled scores can be calculated for all subtests and Cognitive and Social Emotional Scales. Composite scores, percentile ranks and confidence intervals can be calculated for all five scales. Age equivalents and growth scores are available for the Cognitive scale, Expressive and Receptive Language subtests and Fine and Gross Motor subtests. It is important to note that the Technical Manual cautions against the use of age equivalent scores as they are commonly misinterpreted and have psychometric limitations. Bayley (2006) also states that scores on the Bayley-III should never be used as the sole criteria for diagnostic classification (Technical Manual, pg. 84). Scores for the Cognitive, Language and Motor Scales are provided in 10-day increments for children aged 16 days to 5 months and 15 days and in one-month intervals for children over 5 months 15 days. Scaled scores for the Social-Emotional Scale are reported according to the stages of social-emotional development according to Greenspan (2004). Scaled scores for the Adaptive Behavior Scale are reported in 1-month intervals for 0-11 months, 2-month interval for 12-23 months, and 3-month intervals for 24-42 months.

Total administration time ranges between 50 minutes for children younger than 12 months up to 90 minutes for children 13 months and older.

According to the Technical Manual, diagnosing developmental delay can be based on any one of several different criteria: 25% delay in functioning when compared to same age peers; 1.5 standard deviation units below the mean of the reference standard; performance of a certain number of months below the child’s chronological age.

According to the test manual, the Bayley-III may only be administered by trained professionals who have experience in the administration and interpretation of comprehensive developmental assessments and should have completed some formal graduate or professional training in individual assessment. The Bayley-III should not be used to diagnose a specific disorder in any one area. Rather, poor performance in an area should be used to make recommendations for appropriate services.

1. **STANDARDIZATION SAMPLE**

The standardization sample for the Bayley-III included 1700 children aged 16 days through 43 months 15 days divided into 17 age groups each containing 100 participants. Standardization age groups were in 1-month intervals between 1 and 6 months of age, in 2-month intervals between 6 and 12 months of age, in 3-month intervals between 12 and 30 months of age, and in 6-month intervals between 30 and 42 months of age. The standardization sample was collected in the United States between January and October 2004 to match the 2000 United States census*.* The sample was stratified by demographic factors including age, sex, race/ethnicity, geographic region, and primary-care giver’s highest education level. Children were recruited from health clinics, child development centers, speech therapy clinics, hospitals and churches as well as other organizations where children of the appropriate age would be found if they were identified as typically developing and met specific inclusion criteria. A typically developing (TD) child was defined as “any child born without significant medical complications, did not have a history of medical complications, and was not currently diagnosed with or receiving treatment for mental, physical or behavioral difficulties” (Technical Manual pg. 34). Children were excluded if they had confounding conditions or developmental risk factors, were receiving Early Childhood Intervention (ECI) services, did not speak or understand English, did not have normal hearing or vision, and were taking medications that could affect performance or were admitted to hospital at the time of testing. Approximately 10% of the standardization sample included children selected from the special group studies with clinical diagnoses (e.g. Down Syndrome, cerebral palsy, pervasive developmental disorder, premature birth, specific language impairment, prenatal alcohol exposure, asphyxiation at birth, small for gestational age and at risk for developmental delay) to ensure a representative sample. According to the technical manual, these groups were chosen to “more accurately represent the general population of infants and children” (pg. 34). However, according to Pena, Spaulding & Plante (2006), inclusion of children with disabilities in the normative sample can negatively impact the test’s discriminant accuracy, or ability to differentiate between typically developing and disordered children. Specifically, inclusion of individuals with disabilities in the normative sample lowers the mean score, which limits the tests ability to diagnose children with mild disabilities.

The Social-Emotional Scale is based on The *Greenspan* (Greenspan, 2004). In spring 2003 a sample of 456 children aged 15 days to 42 months who matched the U.S. 2000 census were administered The *Greenspan* to generate normative data for the Social-Emotional Scale. The sample was stratified according to parent education level, race/ethnicity and geographic region. The sample was divided into eight age groups each containing a minimum of 50 participants. No mention is made regarding how these children were selected or what percentage, if any of them had clinical diagnoses. These sample sizes are too small according to the standards in the field, which recommends sample sizes of 100 or more (APA, 1974). If a small sample is used then the norms are likely to be less stable and less representative (McCauley and Swisher, 1984).

The Adaptive Behavior Scale is based on the ABAS-II. The standardization sample consisted of 1350 children aged 0-71 months. Children were recruited by telephone calls and flyers from schools, churches and various community organizations. Data were collected between November 2001 and October 2002. Standardization data was collected by 214 independent examiners in 184 cities. The sample was stratified by age, sex, parent education level, race/ethnicity and geographic location. Approximately 2.88% of the sample consisted of children with specific diagnoses including: biological risk factors, language disorders, PDD-NOS, developmental delay, motor impairment, autism and mental retardation (Technical Manual, pg. 50). According the technical manual, children with these specific clinical diagnoses were included in the normative sample according to the percentages reported by the U.S. Department of Education and the theoretical distribution. As mentioned previously, it can be problematic to include children with disabilities in the normative sample as this can negatively impact the test’s discriminant accuracy. The Technical Manual does not explain why one standardization sample was not used to standardize the entire test nor does it acknowledge why a different percentage of children with specific diagnoses (10% and 2.88%) were used for different test components. This variation makes it difficult to compare scores between the three components of the test and reduces the test’s identification accuracy.

1. **VALIDITY**

**Content -** Content Validityis how representative the test items are of the content that is being assessed (Paul, 2007).  Content validity was analyzed via comprehensive literature review, expert review and response processes, which included children’s responses as well as examiner’s observations and interpretations of behavior and/or scores. Test development occurred in a number of phases and revisions were made regularly to ensure appropriate content coverage. The phases included literature reviews, market research, focus groups, semi-structured surveys with experts in child development and a pilot study.

The pilot study consisted of 353 typically developing children as well as two clinical groups (children born prematurely and children with developmental delay). No mention is made regarding demographic information, how these children were selected or number of children in the clinical groups. It is unknown how these children were identified as typically developing or having developmental delay, or if children born prematurely would score differently than the typically developing children.

Following pilot testing, a national tryout phase was conducted using a sample of 1,923 children stratified by demographic variables such as race/ethnicity, caregiver education level, geographic region and sex according to the U.S. 2000 census. An additional sample of 120 African American and Hispanic children were tested to conduct an empirical bias analysis and to ensure adequate sample sizes for the bias analysis. The sole criteria provided for the bias analysis is race. No other information is provided about these children and this is a small sample size. Therefore, it is unlikely these children are a representative sample with which to conduct a bias analysis. Data was also collected from groups of children at risk for clinical disorders (e.g. genetic abnormalities, exposure to toxic substances, attachment disorder, premature birth, chronic disease etc.). Again, it is unclear how the children from the clinical samples were selected, unknown if they score differently than typically developing children and unclear why only children with certain diagnoses were included in the sample. Therefore, we cannot be sure these samples are accurately representative of their groups.

According to the Technical Manual, test items were analyzed during each phase by experts in cross-cultural research and/or child development to ensure content relevance and prevent cultural bias. No mention is made regarding how the test author specifically sought to limit linguistic biases. Items were updated and reorganized and administration and scoring procedures were simplified to ensure content appropriately reflected the construct being measured. Items that were considered biased, redundant or difficult to score/administer were deleted or modified. However, specific information regarding the background and training of the expert reviewers was not provided. 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 cultural biases for which they were evaluating content. As well, no attempt was made by the test author to limit linguistic biases. Therefore, we cannot be certain that test items are free from cultural and linguistic biases.

Content validity was assessed through a variety of methods in order to reduce test biases and increase the clinical utility of the test. However, problems with these methods call into question the content validity of the test. Little information is provided regarding method of selection of sample populations, diagnosis of the clinical populations or sample size. Therefore, we cannot be sure the samples are adequately representative of their groups. As well, no information is provided regarding the training and background of “the expert” panel so we cannot be certain that they were able to adequately assess test content for bias. Therefore, content validity of the Bayley-III is not sufficient.

**Construct –** Construct validity assesses if the test measures what it purports to measure (Paul, 2007).  Construct validity was measured via a series of special groups studies. Inclusion in a group was determined by a “previous diagnosis”, but no information is provided to describe the standards by which these diagnoses were made. The groups were not selected randomly, and the test authors caution that the samples are not completely representative of the diagnostic group because inclusion into each group was not based on defined diagnostic criteria. In addition, according to the Technical Manual, construct validity of the Bayley III “could come from many different sources, including factor analysis, expert review, multitrait-multimethod studies and clinical investigations” (p. 69). Vance and Plante (1994) argue that consulting a panel of experts is apreliminary step in evaluating content validity only. It is not a sufficient way of establishing construct validity (Gray, Plante, Vance, & Henrishsen 1999; Messick 1989). Attempts to determine construct validity for the Bayley-III are not sufficient compared to the standards of the field as determined by Vance and Plante (1994). Due to the lack of information or consistency with regards to how clinical groups were selected, construct validity is not sufficient.

***Reference Standard***

In considering the diagnostic accuracy of an index measure such as the Bayley-III, it is important to compare the child’s diagnostic status (affected or unaffected) with their status as determined by another valid 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).

No reference standard was applied to any of the special groups. Inclusion into each group was based on a previous diagnosis through an unidentified or objective measure. This does not meet the standards set forth by Dollaghan (2007) who states that a reference standard must be applied to both groups, in order to determine the test’s discriminant accuracy. According to Dollaghan (2007), “the reference standard and the index measure both need to be described clearly enough that an experienced clinician can understand their differences and similarities and can envision applying them” (p. 85). The Technical Manual mentions a series of “special group studies” with children who were previously diagnosed with conditions such as Down Syndrome, PDD-NOS, Cerebral Palsy, specific language impairment, at risk for developmental delay, asphyxiation at birth, prenatal alcohol exposure, small for gestational age and premature or low birth weight. These studies elicited scores from children with certain conditions and then compared those scores to a control group. The criteria for inclusion in the control group (typically developing) are not included. No reference standard was used, the special groups are “not representative of the diagnostic category as a whole” (pg. 84) and the sample sizes for each group were very small. Therefore, the reference standard used by the Bayley-III is considered insufficient (Dollaghan, 2007).

***Sensitivity and Specificity***

Sensitivity measures the proportion of students who have a language disorder that will be accurately identified as such on the assessment (Dollaghan, 2007). For example, sensitivity means that Johnny, an eight-year-old boy previously diagnosed with a language disorder, will score within the limits to be identified as having a language disorder on this assessment. Specificity measures the proportion of students who are typically developing that will be accurately identified as such on the assessment (Dollaghan, 2007). For example, specificity means that Peter, an eight-year-old boy with no history of a language disorder, will score within normal limits on the assessment.

The Bayley-III does not provide sensitivity and specificity measure and therefore, cannot be compared to the Vance and Plante standards. Additionally, since the children in the clinical groups were not administered a reference standard their diagnostic status cannot be confirmed. As a result of the lack of sensitivity and specificity information and an unknown reference standard, the diagnostic accuracy of the Bayley-III is unknown. According to Spaulding, Plante and Farinella (2006), a test that does not provide sensitivity and specificity measures should not be used to identify children with language impairment.

***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 Bayley-III could not be calculated because sensitivity and specificity values were not provided.

Overall, construct validity for the Bayley-III was determined to be insufficient for several reasons. The test manual did not describe how typically developing versus delayed/disordered children were distinguished for the normative sample, meaning there was no reference standard. Sensitivity and specificity values were not reported so the accuracy of the test in identifying language disorder/developmental delay is unknown. . Therefore, the Bayley-III cannot be considered a valid test for identifying language disorder/developmental delay and should not be used to identify children with language impairment (Spaulding, Plante and Farinella, 2006).

**Concurrent -** Concurrent Validityis 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 comparison of children’s scores on the Bayley-III and the BSID-II, WPPSI-III, PLS-4, PDMS-2 and ABAS-II Parent/Primary Caregiver Form (ABAS-II-P). Correlations with any of these tests were only moderate at best and of importance, these comparison tests are not valid tests. 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).

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. The Bayley-III was not strongly correlated to any of the comparison tests. Further, many of the tests used are not themselves valid tests either because of unacceptable sensitivity and specificity values (PLS-4 or lack of this information (BSID-II)*.* As well, small sample sizes were examined and most of the tests (PLS-4, WPPSI-III, PDMS-2 and ABAS-II) did not match age ranges to the Bayley-III. None of these tests provided comparison for children below 5 months of age. Therefore, concurrent validity of the Bayley-III is insufficient. This seriously limits the tests ability to correctly answer questions regarding the presence of a language disorder.

1. **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 3 oranges and they weighed 1 pound. If she weighed the same 3 oranges multiple times, and each time they weighed one pound, the scale would have *test-retest reliability*. If other consumers in the store put the same 3 oranges on the scale and they still weighed 1 pound, the scale would have *inter-examiner reliability.* Now say an official were to put a 1 pound calibrated weight on the scale and it weighed 2 pounds. The scale is not measuring what it purports to measure—it is not valid. Therefore, even if the reliability appears to be sufficient as compared to the standards in the field, if it is not valid it is still not appropriate to use in assessment and diagnosis of language disorder.

**Test-Retest Reliability –** Test-retest reliability is a measure used to represent how stable a test score is over time (McCauley & Swisher, 1984). This means that despite the test being administered several times, the results are similar for the same individual. Test-retest reliability was calculated by administering the Cognitive, Language and Motor Scales twice to a group of 197 children. The sample was randomly drawn from the standardization sample and ranged in age between 2-42 months, divided into four age groups (2-4 mo., 9-13 mo., 19-26 mo., 33-42 mo.) each containing approximately 50 children. Reliability coefficients ranged between .67-.94, with correlations increasing as age increased. According to Salvia, Ysseldyke, & Bolt (2010) many of these reliability coefficients are insufficient. They recommend a minimum standard of .90 for test reliability when using the test to make educational placement decisions, such as speech and language services. Also, children from all age ranges for which the Bayley-III is designed were not included and the sample sizes for each age band are small. According to the standards in the field, sufficient sample sizes include 100 or more individuals (APA, 1974). If a small sample is used then the norms are likely to be less stable and less representative (McCauley and Swisher, 1984).

Thus, the test-retest reliability for the Cognitive, Motor and Language Scales is considered insufficient due to a insufficient inclusion of all ages, small sample sizes and because almost all correlation coefficients were less than the accepted minimum standard.

Test-Retest reliability was calculated for the Adaptive Behavior Scale by asking parents to rate their child two times using the same form. The sample included 207 children aged 0-35 months, divided into three age groups (0-11 mo., 12-23 mo. and 24-35 mo.). The test interval ranged between 2-5 weeks. Reliability coefficients range between .74-.92. Almost all of these correlation coefficients are insufficient according to the standards in the field (Salvia, Ysseldyke and Bolt, 2010). Large age bands with few children limit the reliability measure, as well. Thus, the test-retest reliability for the Adaptive Behavior Scale is considered insufficient due to a small sample sizes and because most correlation coefficients were less than the accepted minimum standard. Test-retest reliability was not calculated for the Social-Emotional Scale.

**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 not calculated for the Bayley-III and is thus considered insufficient. This seriously limits the reliability of this test because we cannot be sure the test administration does not affect the test results.

The Technical Manual presents inter-rater reliability for the Adaptive Behavior Scale, which is rated by the child’s parents or primary caregiver. The sample included 56 children aged 0 months – 5 years 10 months each rated by two parents. Reliability coefficients ranged between .59-.86. These are unacceptable according to the standards in the field and suggest a high degree of inconsistency between ratings (Salvia, Ysseldyke and Bolt, 2010). Another issue is that As well, this sample does not reflect the entire age range of the Bayley-III. Therefore inter-rater reliability for the Adaptive Behavior Scale is insufficient.

**Inter-item Consistency –** Inter-item consistency assesses whether parts of an assessment are in fact measuring something similar to what the whole assessment claims to measure (Paul, 2007). Inter-item consistency was calculated by using the split-half method for the Cognitive, Language and Motor subtests and composites using two populations. The normative sample and special groups, which consisted of 668 children with a variety of clinical diagnoses or who were at risk for a diagnosis including down syndrome, developmental delay, Cerebral Palsy, PDD, prenatal alcohol exposure, premature birth, low birth-weight, small for gestational age, and asphyxiation at birth. There were between 46-147 individuals in each band and six of the nine age bands had between 46-53 children in the group. Reliability coefficients are provided combining children in the same age range from all special groups. Reliability coefficients for the subtests range from .86 to .91 for the normative sample and from .94 to .96 for the special groups. The reliability coefficients for the language and motor composites were .93 and .92 for the normative sample. Salvia, Ysseldyke, & Bolt (2010) recommend that the minimum standard for test reliability be .90 when using the test to make educational placement decisions, including SLP services. These reliability coefficients meet the standards in the field according to Salvia, Ysseldyke and Bolt (2010). However, due to small sample sizes in some of the age bands and the inclusion of children with a variety of diagnoses into each group, these coefficients are insufficient to determine internal consistency of the Bayley-III.

Internal consistency for the Social-Emotional Scale was determined via evidence for internal consistency for the *Greenspan Social-Emotional Growth Chart*. Coefficients ranged on average between .76-.90. A number of these correlation coefficients are insufficient according to the standards in the field. Therefore, internal consistency of the Social-Emotional Scale of the Bayley-III is insufficient.

Internal consistency for the Adaptive Behavior Scale was determined via evidence for internal consistency of the ABAS-II. The average reliability coefficient was .97, but some were as low as .75. A number of these correlation coefficients are insufficient according to the standards in the field. Therefore, internal consistency of the Adaptive Behavior Scale of the Bayley-III is insufficient. Internal consistency for the Social Emotional Scale and the Adaptive Behavior Scale was determined through internal consistency for the scales on which they are based. No study was conducted to assess reliability of these scales themselves as parts of the Bayley-III. Therefore, even though some correlation coefficients meet acceptable standards in the field reliability cannot be considered sufficient, as it does not adequately reflect the Social emotional scale and Adaptive Behavior Scale scale/Questionnaire.

**STANDARD ERROR OF MEASUREMENT**

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

The Bayley-III provides SEM for the Cognitive, Motor and Language Subscales and Composites. SEMs for the subscales range between .93 – 1.19 (based on SD of 3) and SEM’s for the composites are 4.47 and 4.42 (based on SD of 15). Average SEM for the Social-Emotional Scale is 1.00 and for the Adaptive Behavior Scale General Adaptive Composite the SEM is 3.11. The Bayley-III uses the SEM to determine the confidence intervals for the Adaptive Behavior Composite (GAC). The Bayley-III uses the Standard Error of Estimation (SEE) to determine confidence intervals for the Cognitive, Social Emotional, Language and Motor composite scores. According to the Technical Manual, confidence intervals around the SEE and SEM can be interpreted the same way and are used to calculate the band of scores within which an individual’s true score is likely to fall (pg. 112). The use of SEE is a method to account for true score regression to the mean. However, the manual does not explain why this is beneficial over the use of SEM (Technical Manual, pg. 60). The manual encourages the use of confidence intervals to ensure more accurate interpretation of scores.

The clinician chooses a confidence level (usually 90% or 95%) at which to calculate the confidence interval. A higher confidence level will yield a larger range of possible test scores, including a child’s true range of possible scores. Although a larger range is yielded with a higher confidence interval, the clinician can be more *confident* that the child’s ‘true’ score falls within that range. A lower level of confidence will produce a smaller range of scores but the clinician will be less confident that the child’s true score falls within that range. The wide range of scores necessary to achieve a high level of confidence, often covering two or more standard deviations, demonstrates how little information is gained by administration of a standardized test.

The Bayley-III provides confidence levels at 90% and 95% that are based on the SEE rather than SEM. Consider a 24 month-old child who obtained a language composite score of 79. As this score falls 1.5 SD from the mean, this child would be classified as having a mild language disorder and given services, which can have a serious negative effect on his life. However, according to the manual, at a 90% confidence level, the child’s true score falls between 74 and 87. The lower bound of this interval would suggest a moderate to severe 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). 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.

1. **BIAS:**

**Linguistic Bias**

***English as a Second Language***

Paradis (2005) found that children learning English as a Second Language (ESL) may show similar characteristics to children with Specific Language Impairments (SLI) when assessed by language tests that are not valid, reliable, and free of bias. Thus, typically developing students learning English as a Second Language may be diagnosed as having a language disorder when, in reality, they are showing signs of typical second language acquisition. According to ASHA, clinicians working with diverse and bilingual backgrounds must be familiar with how elements of language differences and second language acquisition differ from a true disorder (ASHA, 2004).

The Technical Manual does not make mention of any potential limitations in using the Bayley-III with children with limited English proficiency (LEP). According to Paradis (2005), grammatical morphology has been noted as an area of difficulty for children with LEP. In the Expressive Communication subtest of the Language Scale the child is required to describe pictures using the verb + *ing*. If the child uses the correct verb, but fails to conjugate it into the present progressive tense, this item will be scored as incorrect (Administration Manual, pg. 113). A child who has LEP might have difficulty with this form and respond incorrectly, reflecting the child’s lack of exposure to English rather than a disorder. Despite research demonstrating the similarity in expressive language of children with LEP and children with SLI, the manual makes no mention of this fact to alert clinicians to the potential inappropriateness of the test item.

***Dialectal Variations***

A child’s performance on the Bayley-III may also be affected by the dialect of English that is spoken in their homes and communities.

Consider dialect issues resulting from the test being administered in Standard American English (SAE). For example, imagine being asked to repeat the following sentence, written in Early Modern English: “Whether 'tis nobler in the mind to suffer the slings and arrows of outrageous fortune or to take arms against a sea of troubles and by opposing end them" (Shakespeare, 2007). Although the content of the sentence consists of words in English, because of the unfamiliar structure and semantic meaning of an archaic dialect, it would be difficult for a speaker of SAE to repeat this sentence. Speakers of dialects other than SAE (e.g. African American Vernacular English [AAE], Patois) face a similar challenge when asked to complete tasks that place a large emphasis on and grammar. Dialectal differences between the child and the tester can affect the student’s performance on all subtests. For example, some items on the expressive language subtest of the language scale focus on understanding plurals. Children who speak/have primarily been exposed to AAE will not have the same understanding of plurals compared to children who speak/have primarily been exposed to Standard American English. On the Expressive Communication subtest of the language scale item 38, the child is required to describe a series of pictures that all contain multiple items. A correct answer requires the child to use the plural “s” when labeling the picture. In AAE, the plural “s” is frequently omitted from a word. Therefore a child who speaks AAE may score incorrectly on this item despite understanding the plural concept. Items 46, 47 and 48 s the child is required to “tell a story about this picture” in the past tense. Individuals who speak often emit the past tense or use other words to mark tense. For example, “she has eaten her dinner” could be expressed as “she done eat her dinner”; “she went to sleep” could be expressed as “she had went to sleep” (Wolfram, 2000). On other subtests, the child’s performance could be affected by the examiner’s dialect or his or her understanding of the child’s dialect. Differences in expressive language limit the child and tester’s ability to understand one another and the child’s performance can be misinterpreted (Delpit, 1995). 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.

**Socioeconomic Status Bias**

Hart & Risley (1995) found that a child’s vocabulary correlates with his/her family’s socio-economic status; parents with low SES (working class, welfare) used fewer words per hour when speaking to their children than parents with professional skills and higher SES. Thus, children from families with a higher SES will likely have larger vocabularies and thus will likely show a higher performance on standardized child language tests.

Many of the questions in the expressive and receptive language subtests from the Language scale require a child to describe pictures and are an assessment of vocabulary skills. Hart & Risley (1995) report that children from welfare backgrounds are exposed to an average of one-third the number of words that children from a professional working background are exposed to and half the number of words that children from a working class background are exposed to. This opportunity to learn more vocabulary gives children from higher SES groups an advantage on the assessment. Children from low SES consistently 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).

**Prior Knowledge/Experience**

A child’s performance on Bayley-III may also be affected by their prior knowledge and experiences. For example, a child from a large city may not know that a frog does not have a tail (item 48, Receptive Language Scale). Or a child from an urban center might not know that a hammer’s handle is made out of wood (item 47, Receptive Language Scale). A correct answer to these questions may depend on the child’s prior experience with the item. Therefore, an incorrect answer may reflect different experience rather than receptive language deficit.

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. There are a number of items on the Cognitive and Language Scales that assess whether a child can answer questions logically and explain how an object is used. In some cultures it is inappropriate to ask known questions. Children taking the assessment may be confused why the examiner is breaking this cultural norm, affecting performance.

Children who have prior knowledge and experiences with test content and testing environment are at an advantage. For example, in the Expressive Communication Subtest, describing picture series may favor students who have prior exposure to stories and is biased against children from low SES have less familiarity with stories (Hart & Risley, 1995).

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. On the Bayley-III the child is required to attend to pictures in the stimulus book in order to complete items in the Cognitive and Language Scales. Many items on the cognitive scale require the child to attend to and manipulate puzzles and books, as well. For example, item 32 on the Cognitive Scale requires the child to look at pictures from a small picture book. A series of items require the child to manipulate pegs in a pegboard or put puzzle pieces together. Lack of experience or exposure to these kinds of objects and books may negatively affect a child’s performance.

**Cultural Bias**

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

It is also important to consider that during test administration, children are expected to interact with strangers. In middle class mainstream American culture, young children are expected to converse with unfamiliar adults as well as ask questions. In other cultures, however, it may be customary for a child to not speak until spoken to. When a child 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. The format of much of the Bayley-III requires the child to respond to verbal prompts provided by the examiner. Peña and Quinn (1997) observed that while reading together, European American mother-child dyads appeared to be in more of a question/answer mode compared to African American mother-child dyads. Children who regularly interact with adults in question/answer mode have task familiarity and will perform better on assessments of this nature.

Social factors influence the way children learn to speak and interact. According to Heath (1982), language socialization occurs as the child grows up and affects how they learn or “take” information from their environment, how they talk in social interactions and how they express information. Children may feel uncomfortable or be unfamiliar with how to demonstrate their abilities to an adult for social and cultural reasons and the examiner must consider what the child has been exposed to in their environment when assessing that child.

**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.

Bayley-III administration time is approximately 50 minutes for children younger than 12 mo., and approximately 90 minutes for children older than 13 mos. This can be very taxing for young children particularly if they are uncomfortable with the testing environment or if the material is challenging. The Administration Manual encourages testers to adapt their testing techniques to the child’s needs, temperament and disposition. All subtests must be administered in one session, but if the child becomes very fatigued or restless the examiner can “allow the child to take a 5-minute break or have a snack if needed” (Administration Manual, pg. 15). This is helpful, but not a significant amount of time to really help the child if he/she is exhausted from the testing. This can be especially difficult for very young children as the Bayley-III can be administered to children as young as 16 days.

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.

**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.

The Bayley-III is designed to assess developmental delay in a number of areas of development including motor and sensory skills. Further, the sample population included children with motor and sensory impairments. Specific information regarding the nature or extent of their motor and sensory impairments is not provided. The Bayley-III should be not be used to measure the nature of the deficit in each specific area (Technical Manual, pg. 84). The Bayley-III explicitly states that test administration should be modified to accommodate children with specific physical or language impairments (e.g. sign language, visual aids or hearing impairment). However, if modifications are made they will invalidate the use of normative scores (Administration Manual, pg. 45).

1. **SPECIAL ALERTS/COMMENTS**

*The Bayley Scale of Infant and Toddler Development (Bayley-III) is designed to assess the developmental functioning of infants and young children 1-42 months of age. The Bayley-III attempts to identify suspected developmental delays in children and to provide information to plan and develop appropriate interventions rooted in theory and research in the area of child development. However, results obtained from administration of the Bayley-III are not valid due to lack of an adequate reference standard and lack of sensitivity and specificity data. This seriously limits the test’s discriminant accuracy and ability to properly identify children with a developmental delay or disorder. Furthermore, there are other issues of concern that call into question the validity and reliability of the Bayley-III. The test contains significant cultural and linguistic biases, which preclude it from being appropriate for children from diverse backgrounds. Even for children from mainstream, SAE speaking backgrounds, the test has not demonstrated adequate validity and diagnostic accuracy. According to federal**legislation, testing materials are required to be “valid, reliable and free of significant bias”(IDEA, 2004).*

*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. The test will likely falsely identify children from non-mainstream cultural and linguistic backgrounds or who come from lower socioeconomic status as language delayed or disordered. 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 & Klingner, 2006) and higher dropout rates (Hehir, 2005). Therefore, scores should not be calculated or used to diagnose speech and/or language delay/disorder or to determine special education services. A speech and language evaluation should be based on clinical observations, including clinical judgment, consideration of the child’s prior experiences and development history and parent 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.

Gray S, Plante E, Vance R, & Henrichsen M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*, 196–206.

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

Horton-Ikard, R., & Weismer, S. E. (2007). A preliminary examination of vocabulary and word learning in African American toddlers from middle and low income socioeconomic status homes. *American Journal of Speech-Language Pathology, 16*, 281-392.

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.

Messick S.(1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5–11.

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.

Wolfram, W. (2000). The grammar of urban African American Vernacular English*.* In Eds. B. Kortmann and E. Schneider. *Handbook Varieties of English* (111-132). Berlin: Mouton de Gruyter.