

Running head: CONVENTIONAL TESTING FOR EARLY INTERVENTION ELIGIBILITY

Research Foundations of Conventional Tests and Testing to Ensure Accurate and Representative  
Early Intervention Eligibility

Marisa Macy

Penn State

Stephen J. Bagnato

Cathryn Lehman

Jen Salaway

Children's Hospital of Pittsburgh/  
University of Pittsburgh

Bagnato, S.J., Macey, M., Salaway, J., & Lehman, C. (2007). Research foundations for conventional tests and testing to ensure accurate and representative early intervention eligibility. Pittsburgh, PA: TRACE Center for Excellence in Early Childhood Assessment, Early Childhood Partnerships, Children's Hospital/University of Pittsburgh; US Department of Education, Office of Special Education Programs, and Orelena Hawks Puckett Institute.

## Abstract

Conventional norm-referenced tests, which are nationally standardized, are designed to estimate a child's level of functioning under a controlled and preset series of conditions. Conventional tests of early development and psycho-educational functioning are often used to determine eligibility for early intervention and early childhood special education programs and services. For the norms to be scientifically accurate for this purpose, the examiner must use standardized methods and instructions (Newborg, 2005). However, conventional tests and testing practices must have salient attributes to ensure accurate and representative measurement of the capabilities of infants, toddlers, and preschool children who have delays and disabilities.

In this research synthesis, we identified 6 characteristics of conventional tests and testing practices that are necessary for use with young children to determine their eligibility for early intervention services; we, then, applied these test characteristics to nine commonly used conventional tests/editions. Next, we reviewed the available research on the most frequently used conventional tests and found 29 studies. Findings of this synthesis will help professionals to critically identify characteristics of conventional tests and testing practices that influence the accurate and representative documentation of a young child's degree and pattern of delay or disability to determine early intervention eligibility.

## Research Foundations of Conventional Tests and Testing to Ensure Accurate and Representative Early Intervention Eligibility

In the United States, young children are frequently tested to determine if they are eligible to receive early intervention (EI) and early childhood special education (ECSE) services as defined in the Individuals with Disabilities Education Improvement Act (IDEA, 2004; P.L. 108-446). Part of the eligibility determination process involves using technically sound measurement to appraise individual competencies expected for age. For the purposes of this synthesis, “conventional” tests refer to standardized and norm-referenced measures of early development and psycho-educational functioning that are commercially available and widely used for determining eligibility for IDEA services.

Common features of conventional tests include a high degree of structure whereby the practice of testing is mostly restricted to professionals with specialized training and qualifications that must follow strict testing procedures. These normative tests have criteria associated with individual test items and therefore show some criterion-referenced attributes. Conventional test items often follow basal/ceiling rules which have starting and end points that are both age and performance-based. Standardized materials and procedures are typically used during conventional testing to allow for group comparisons. The administration format most often includes direct testing which is used to elicit specific behaviors from an individual child. Conventional tests are frequently administered in a clinical setting or under defined, laboratory-like conditions.

There are several advantages to using conventional assessment. First, conventional tests can provide a diagnosis, description, eligibility determination and/or prediction of future performance (Bagnato & Neisworth, 1994). Conventional test scores are generally used to

determine if a child meets the eligibility criteria for his or her state (Bailey & Wolery, 1989; McLean, Bailey, & Wolery, 2004). Therefore, conventional tests answer the question, “Is this child eligible for early intervention services?” Second, conventional tests supply normative data (Kelly-Vance, Needelman, Troia, & Ryalls, 1999). Conventional tests are meant to differentiate those who can and cannot perform specific skills. Most conventional tests are normed on a representative sample of typically-developing peers (norm-referenced). Children are compared with a normative sample, and their relative placement is expressed in a standard score format that enables a stable comparison across ages and tests (Wilson, 1983). Third, conventional tests have established reliability and validity (McLean, Wolery, & Bailey, 2004; Sattler, 2001). Conventional testing provides reliable and valid quantitative information for eligibility determination through an appraisal of a child’s mental development (Magiati & Howlin, 2001). These data are usually reported in test manuals and available to consumers. Psychometric data are also essential for research purposes, when it is necessary to determine baseline levels of ability, to ensure that treatment and comparison groups are closely matched.

Despite the apparent advantages, conventional tests and testing practices have been criticized for not having attributes that ensure an accurate and representative appraisal of the capabilities of young children with delays and disabilities. Some of the neglected attributes include procedural flexibility to adapt testing materials and formats to accommodate the child’s functional impairments; inclusion of children in the standardization sample who have a range of diverse disabilities; a sufficient density of items and content to enable the appraisal of low functional levels; graduated scoring options which enable a more finely-graduated appraisal of the child’s full-range of capabilities; and inclusion of child performance data across several people, settings, and occasions.

Despite both the advantages and limitations, state agencies have produced regulations that make conventional tests and testing preferred choices to test young children in order to make decisions about eligibility for IDEA services. *Infants and toddlers* who are identified as having special educational and developmental needs, and qualify for early intervention (EI), are served under Part C of IDEA; while *preschoolers* who qualify for early childhood special education (ECSE) programs are served under Section 619 of Part B. Each state is required to define their eligibility requirements (appropriate instruments and procedures) by measuring the levels of functioning or other criteria and procedures it will use to determine the existence of a delay in the following areas of development: cognitive, physical (including vision and hearing), language and speech, social/emotional, and adaptive. There are variations among state policies in terminology, age range assigned to the category, eligibility criteria and restriction on its use (Danaher, 2005). For Part C for example, 48 states and jurisdictions out of 58 have the requirements of demonstrating a standard deviation, percentage of delay, delay in months, or percentile scores that can be obtained by a norm-referenced, conventional assessment (Shackelford, 2006). In Part B, 43 states use quantitative criteria for determining developmental delay; 38 states or jurisdictions use standard deviation and 17 use percent delay (Danaher, 2005). Part B and C regulations permit specific scores to provide a standard deviation, percentage of delay, and age equivalency for determining eligibility for IDEA services.

### Purpose

There are two purposes for this research synthesis: (1) to identify *characteristics of conventional tests and associated testing practices* that are necessary to document a child's degree of delay and disability in an accurate and representative manner in order to qualify children for EI and ECSE services; and (2) to synthesize the body of *research on specific*

*conventional tests* of early development and psycho-educational functioning and to appraise their capacity to determine IDEA eligibility for EI/ECSE services in an accurate and representative manner. This practice-based research synthesis examines available evidence regarding the following five conventional tests:

1. Battelle Developmental Inventory (BDI)
  - a. BDI-1; Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984
  - b. BDI-2; Newborg, 2005
2. Bayley Scales of Infant and Toddler Development (BSID)
  - a. BSID-II; Bayley, 1993
  - b. BSID-III; Bayley, 2006
3. Mullen Scales of Early Learning, AGS Edition (MSEL; Mullen, 1995)
4. Stanford Binet Intelligence Scales (SB)
  - a. SB: IV; Thorndike, Hagen, & Sattler, 1986
  - b. SB5; Roid, 2006
5. Wechsler Preschool & Primary Scales of Intelligence
  - a. WPPSI-R; Wechsler, 1989
  - b. WPPSI-III; Wechsler, 2002

These five conventional tests were chosen through a review of literature and surveys of practitioners investigating the most commonly used measures in preschool and early intervention settings (Bagnato & Neisworth, 1994; Bradley-Johnson, 2001; Flanagan & Alfonso, 1995; Goh, Teslow, & Fuller, 1981; Hutton, Dubes, & Muir, 1992; Pretti-Frontczak, Kowalski, & Douglas-Brown, 2002; Mardell-Czudnowski, 1996; Wilson & Reschly, 1996).

*Test and Testing-Associated Practice Characteristics*

Tests and testing practices used for the purposes of determining a child's eligibility for early intervention should meet specific standards. After an initial review of the measurement literature in psychology, special education, early intervention, and early childhood special education, we identified six characteristics that represent quality indicators for conventional tests and testing practices. Procedures used to determine eligibility for EI/ECSE should contain the following characteristics: (1) disability samples, (2) procedural flexibility, (3) comprehensive coverage, (4) functional content, (5) item density, and (6) graduated scoring.

*Disability sample* refers to the importance of ensuring that the standardization included young children with delays/disabilities in the normative group and any field-validation samples. Research indicates that it is critical that children with functional characteristics similar to the child being tested to be included in the standardization/field-validation sample. The sample should have at least 100 people per each age interval.

The second test or testing-associated practice characteristic is *procedural flexibility*. Procedural flexibility refers to the extent to which the test administration procedures allow professionals to modify the method of testing (i.e., table-top vs. play), the stimulus attributes of items, and the response modes of the young child to accommodate their functional impairments leading to a more realistic and representative estimate of capabilities.

Conventional tests and testing procedures must ensure *comprehensive coverage* of multiple domains of developmental functioning. Such broad coverage generates results that profile the young child's capabilities across multiple and interrelated functional competencies (i.e., cognitive, motor, adaptive, communication, self-regulatory). Simply, the results of eligibility testing must reflect the "whole child."

*Functional content* indicates that test items should be comprised of content that requires functional and meaningful skills for everyday life rather than discrete and isolated tasks.

Conventional tests have sufficient *item density* when enough tasks or content are included in the survey of each developmental domain so that even low functional levels can be profiled--the lowest range of standard scores can be obtained when a child does not pass test items, or only a few test items were scored correctly on a test or subtest. The test items should be low enough to discriminate age (i.e., very young children) and level of functioning. There should also be a comprehensive set of items in each age interval to describe a child's performance.

The conventional tests should encompass a *graduated scoring* system to reflect varied skill levels on individual items. For example, a three-point scoring rubric (e.g., yes, emerging, and not yet) provides more information about a child's development rather than a simple two-point scale (e.g., yes/no), and may also indicate the conditions under which a young child can or cannot perform (i.e., with physical prompts, with verbal prompts, with general assistance, independently). The practice characteristics were evaluated by gathering information from test manuals.

#### *Research-based Use Characteristics*

Research studies on the identified conventional tests were reviewed in this synthesis. More specifically, we examined *research characteristics* associated with the **use** or application of each test. We examined tests regarding how accurate, representative, and reliable they were in generating results applicable to eligibility determination. The studies were reviewed to provide support for specific "use" or applied effectiveness characteristics determined to be significant when using conventional tests to establish early intervention eligibility.

Research characteristics, determined by data synthesis and professional judgment, include *accuracy* (reliability) and *effectiveness* (validity, utility). Accuracy refers to the extent to which a tool identifies young children with disabilities, specifically. This includes the reliability of the tool, which includes consistency across test items and the use of cut-off scores in order for the tool to precisely or accurately measure a skill or behavior. Examples of accuracy include: test retest reliability, inter-rater reliability, intra-rater reliability, and inter-item consistency.

Effectiveness refers to the extent to which a tool successfully identifies young children with disabilities. This includes the validity of the measure (i.e., including to what extent does the tool measure what it was designed to measure) and how it relates significantly to similar measures. Examples of effectiveness include: predictive validity, concurrent validity, construct validity, test floors, and item gradients.

### Search Strategy

#### *Search Terms*

Relevant published literature and unpublished position papers, literature reviews, and research studies were identified using the following search terms: assessment (conventional, traditional), testing, early intervention, preschool, early childhood, eligibility, pediatrics, disabilities, handicap identification, referral, preschool children specific assessment tool were searched. More general terms of special schools, state programs resource, centers and evaluation were also used.

The five conventional tests were also included within the search. The search was done broadly in the fields of psychology, developmental disabilities, special education, allied health fields (speech and language therapy, physical therapy, occupational therapy), as well as early intervention.

*Sources*

The primary databases included the following sources: CINAHL, Cochrane Library, Digital Dissertations, Ebsco Host, Education Resource Information Center (ERIC), Google Scholar, Health Source, Illumina, Ingenda, Inter Science, Medline, Medscape, Ovid, Psychological Abstracts (PsycINFO), Social Sciences Citation Index, and Springerlink. Additionally, we conducted selective searches of unpublished master's theses and doctoral dissertations. Hand searches of select journals and ancestral searches were also conducted.

*Selection Criteria*

Efficacy for eligibility determination based on the practice and research-based use characteristics was examined using five conventional tests: BDI (Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984; Newborg, 2005), BSID (Bayley, 1993; 2006), MSEL (Mullen, 1995), SB (Thorndike, Hagen, & Sattler, 1986; Roid, 2006), and WPPSI (Wechsler, 1989, 2002). The literature review and synthesis was based upon over 300 articles. The study had to meet the following criteria for inclusion: (a) researched one or more of the popular conventional tests, (b) involved the evaluation of young children with disabilities or at-risk for developing a disability due to environmental or biological risk conditions, (c) examined the accuracy of the measure at testing infants, toddlers, and preschool children with disabilities, and (d) was published in a scientific and scholarly publication. This overall synthesis was conducted as part of literature reviews and syntheses conducted at the Tracking, Referral and Assessment Center for Excellence (Dunst, Trivette, & Cutspe, 2002).

## Results

### *Conventional Test and Testing-Associated Characteristics*

A review of the literature and guidelines in Standards for Educational and Psychological Testing (AERA, 1999) provided the basis for the development of test characteristics. The purpose of Table 1 is to show test characteristics that are needed to determine young children eligible for early intervention. Characteristics were: (1) disability sample, (2) procedural flexibility, (3) comprehensive coverage, (4) functional content, (5) item density, and (6) graduated scoring. We examined the BDI, BSID, MSEL, SB, and WPPSI tests for these six characteristics. All of the tests, except the MSEL, had more than one edition for a total of nine tests.

Of the nine tests, only two (22%) included children with disabilities in the standardization sample. There were no tests that had procedural flexibility (0%) or item density (0%). However, the BDI-1 and BDI-2 permit test accommodations for children with disabilities. Four tests (44%) had comprehensive coverage to include multiple developmental domains. There were only two tests that included some functional test items (22%), as well as graduated scoring options (22%).

- **ALL** of the necessary test characteristics were missing in over half of the tests.
- Only 4 of the conventional tests had some quality characteristics.
- The following tests were missing **ALL** of the six quality characteristics: BSID-II, SBIV, SB5, WPPSI-R, and WPPSI-III.
- The BDI-2 had the most quality characteristics of the nine tests.

<insert Table 1 here>

Information about each of the nine tests can be found in Table 2. Since most of these conventional tests had multiple editions, a side-by-side comparison of these measures is provided to include: (1) developmental domains, (2) developmental sub-domains, (3) number of items, (4) age range, (5) normative sample, (6) children with disabilities included in the standardization, (7) norm intervals, (8) test accommodations allowed for children with special needs, and (9) test citations. The BDI-2 (2005), BSID-III (2006), and SB5 (2006) are updated editions that have been published recently; therefore they are too new to have a substantial empirical base yet.

- All conventional tests, except the BDI, prohibited the use of testing accommodations for children with special needs.
- Only 2 recently published conventional tests (i.e., BDI-2 and BSID-III) included children with disabilities in the standardization sample. <insert Table 2 here>

#### *Research Characteristics Related to Five Conventional Tests*

A total of 29 studies on conventional tests were identified from the fields of child development, early intervention, psychology, special education, physical therapy, pediatrics, and behavioral development. The most studies were found on the BDI and BSID, whereas the MSEL and WPPSI only had a few studies that met our search criteria. Here we report the: total number of studies that met the inclusion criteria, years articles were published, age range included in the studies, and the total number of participants in study samples for the BDI, BSID, MSEL, SB, and WPPSI.

#### BDI

- 16 studies
- published between 1984 and 2000
- age range was birth to 95 months
- 1,637 young children
- *Note* - none of the studies used the BDI-2 (2005) version

BSID

- 11 studies
- published between 1985 and 2004
- ages ranged from 2 weeks to 77 months
- 1,043 young children
- *Note* – none of the studies used the BSID-III (2006) version

MSEL

- 2 studies
- published between 1995 and 1999
- ages ranged from 0 to 48 months
- 237 young children

SB

- 6 studies
- published between 1992 and 2003
- ages ranged from 18 to 132 months
- 734 young children
- *Note* – none of the studies used the SB5 (2006) version

WPPSI

- 3 studies
- published between 1992 and 2000
- ages ranged from 36 to 72 months
- 450 young children

### *Participants*

Overall, there were 3,150 young children who participated in the 29 studies. Children's ages ranged from birth to 132 months. Twenty studies (69%) included children with identified disabilities or delays, and 12 studies (41%) included children at risk<sup>1</sup>.

Children were identified as developmentally delayed, mentally retarded, speech/language delayed, cerebral palsy, premature birth, Down syndrome, microcephaly, Cri Du Chat syndrome, intraventricular hemorrhage, emotional disturbance, metabolic disorders, visual function categories, utero-cocaine exposure, autism, brain injury, fetal alcohol syndrome, hydrocephalus, and spina bifida. Some studies included children and families from diverse cultural and linguistic backgrounds.

Professionals that performed the testing included clinical psychologists, school psychologists, graduate students in clinical or school psychology, master's level child/pediatric psychology assistants, early childhood special education personnel, teachers, service providers in early intervention, speech and language therapists, and physical therapists. Parent's judgments were included in some studies to complete testing in which conventional tests failed to yield useful results for eligibility determination. In general, the testing was accomplished in a clinical setting, a center-based school setting, or in the child's home. The purpose of Table 3 is to show demographic characteristics of children which include: total sample size, mean age in months, age range in months, and child ability characteristics.

<insert Table 3 here>

---

<sup>1</sup> Some studies included both children with *disabilities* and *at-risk* for developing a disability due to medical (e.g., low birth weight) or environmental (e.g., exposure to illegal drug use, teen parent) conditions. Therefore, the overall percentage of children included in sample exceeds 100%.

### *Types of Studies*

Each study reported in this synthesis examined some aspect of accuracy and/or effectiveness related to one or more of the conventional tests. We found the following types of studies: 5 inter-item/inter-rater reliability, 2 test-retest reliability, 2 sensitivity/specificity, 14 concurrent validity, 7 predictive validity, 3 construct and 1 criterion validity, and 2 utility studies. Accuracy (reliability) and effectiveness (validity) of the 29 research studies are identified in Table 4.

<insert Table 4 here>

### *Outcomes*

*Accuracy (reliability).* Table 5 incorporates results on the accuracy and effectiveness of conventional tests. A total of ten studies ( $n = 10/29$ , 34%) examined the accuracy of conventional tests. There were five studies that looked at how accurate the **BDI** is in identifying children with special needs which included *test-retest reliability* and *inter-rater reliability*. Using five BDI protocols, test-retest resulted in .80 reliability coefficient within six weeks of the first BDI administration (Boyd, Welge, Sexton, & Miller, 1989). Inter-rater reliability coefficients for the BDI were very good with: .97 item-by-item (Boyd et al., 1989), .90 (Hatton, Bailey, Burchinal, & Ferrell, 1997), .93 on all items and above .90 across all domains (McLean, McCormick, Bruder, & Burdg, 1987), .93 item-by-item agreement (Sexton, McLean, Boyd, Thompson, & McCormick, 1988), and above .85 (Snyder, Lawson, Thompson, Stricklin, & Sexton, 1993).

Mixed findings were reported for 3 studies that examined the accuracy of the **BSID**. Studies showed evidence for test-retest reliability, but not sensitivity and specificity. Cook, Holder-Brown, Johnson, and Kilgo (1989) found significant reliability coefficients between the

BSID Mental Scales at 6 months and then again at 12 months. However, Mayes (1999) found that scores on the BSID-II were skewed positively when starting at the item set closest to the child's age, suggesting that the BSID-II lacks sensitivity. Similarly, Kelly-Vance, Needelman, Troia, and Ryalls (1999) found that 82% of a sample of children with low birth weight would not have been identified for EI services based on the results of the BSID-II. Results of this study indicate that the BSID-II is not adequately sensitive in identifying low birth weight children in need of EI services.

The accuracy of the **MSEL** in identifying children with special needs showed it lacked sensitivity and specificity (McConachie, Couter, & Honey, 2005). When testing young children using the MSEL, results indicated that only three children met criteria for Autism Spectrum Disorder (ASD); however, 79 children from the sample were later diagnosed with ASD. This suggests that the MSEL lacks sensitivity and specificity in detecting symptoms of ASD.

In 2000, Grunau, Whitfield, and Petrie examined the accuracy of the **SB**. Specifically, the study found that children with full scale IQs of less than 84 on the **WPPSI-R** at ages 4 and 5 were poorly identified (sensitivity of 54%) from the SB composite at age 3, suggesting that the SB IV is not a sensitive assessment for children with biological risk factors.

*Accuracy related to test characteristics.* The level of *accuracy* related to constructs in the research studies was influenced by the presence or absence of the established test characteristics (i.e., disability sample, procedural flexibility, comprehensive coverage, graduated scoring, functional content, and item density).

Ideally, tests should have a normative sample where people who take the test (i.e., children eligible for early intervention) were included in the standardization process to increase the accuracy of test outcomes. Unfortunately in our synthesis, we found no studies that had a

representative *disability sample* from the list of tests we reported in Table 1. In other words, there were two tests from our list that included children with disabilities; however those two tests were not measures used in the research studies we reviewed possibly because they are recent editions. The absence of children with disabilities in the normative sample makes it difficult to understand the nature and degree of an eligible child's special needs, because we would be measuring that child against a peer group with differing capacities. Sensitivity and specificity studies can provide information about how accurately a conventional test classifies children as eligible or not eligible. Sensitivity refers to a test being able to correctly classify children with delays and/or disabilities who are eligible for early intervention services. Specificity refers to a test being able to correctly classify children with typical development who are not eligible for early intervention services. The less than adequate outcomes of the sensitivity and specificity studies in our review demonstrated a need to better understand the performance of children with disabilities in the test sample (Grunau et al., 2000; Kelly-Vance et al., 1999; Mayes, 1997; McConachie et al., 2005). By including children with disabilities in the normative sample, conventional tests could address issues of test sensitivity and specificity when determining a young child eligible for early intervention services.

*Flexible procedures* are important to accuracy because it allows accommodations for young children with special needs. Test results are more likely to be accurate, because the test examiner will accommodate the child's special needs to determine if the child can or cannot perform a task on the test to avoid measurement error. For example, a child with little or no vision will perform differently on a test compared to children who are sighted; moreover they may need test accommodations. Rigid adherence to a set of procedures could impact the performance of a child with vision loss by restricting his/her responses on tasks that may not be

appropriate or accurate. Unfortunately, of the research studies included in our synthesis, we found no studies that had procedural flexibility. The absence of flexible options when administering conventional tests was a limiting factor that was evident with the poor sensitivity and specificity outcomes for children with ASD (Mayes, 1997; McConachie et al., 2005) and risk factors (Grunau et al., 2000; Kelly-Vance et al., 1999). Tests that are reliable produce similar outcomes across a variety of circumstances – like administration flexibility.

Comprehensive coverage leads to more accurate decisions about eligibility because the test will contain multiple and varied content. The advantage of an eligibility assessment over a screening assessment is that there are more test items that cover a larger span of development. A screener allows the consumer to decide if further assessment is warranted, whereas an eligibility assessment should inform the consumer if the child qualifies for early intervention services and requires more evidence. Therefore, the conventional test that is used to determine eligibility for early intervention must cover the topography of child development needed to make those decisions, and when it has comprehensive coverage is more likely to be consistent. Graduated scoring improves accuracy because it provides a more precise interpretation of skill level. The presence of these two test characteristics, *comprehensive coverage* and *graduated scoring*, showed favorable results on one of the tests because inter-rater reliability ranged from .85 to .97 on four of the studies (Boyd et al., 1989; McLean et al., 1987; Sexton et al., 1988; Snyder et al., 1993). Multiple developmental domains and ability to capture varying levels of performance through a graduated scoring system are recommended test characteristics to improve accuracy by creating a systematic way to show what a child can do and still needs to learn or develop.

Functional content represents meaningful skills needed for the child to become independent in his/her usual settings. Of the research studies we reviewed, one test used in the

research had functional content. Test-retest was found to be in the adequate range (80%) for the test that had functional content (Boyd et al., 1989; Cook et al., 1989). A child is administered a test once and then again within a specific time frame (e.g., 2 weeks later) for a test-retest study. Children's performance on the test with functional content remained consistent over time.

*Item density* is another quality indicator for test accuracy. When an eligibility test is missing this feature, test consistency in identifying infants and toddlers for early intervention is suspect. There were no tests in our review that met this standard for item density. Consistency and reliability improves when there is a density of test items within age intervals.

*Effectiveness (validity)*. A total of 26 studies ( $n = 26/29$ , 90%) examined the efficacy of conventional tests. Concurrent validity was the most common type of research. Fifteen studies examined the effectiveness of the **BDI** by investigating: utility (1 study), concurrent validity (9 studies), construct (3 studies) and criterion validity (1 study), and predictive validity (3 studies).

The BDI was used with teachers who found it to be a useful test for children with mild to moderate delays, but less useful when used with children who have more severe delays (Bailey, Vandiviere, Dellinger, & Munn, 1987). Bailey et al., (1987) also noted that users made many mistakes on the BDI protocols and administration. Studies on the concurrent validity of the BDI were conducted with various other tests and found to have mixed results; however a number of studies showed the BDI had weak correlations with other tests like the BSID (Gerken, Eliason, & Arthur, 1994; Johnson, Cook, & Kullman, 1992) and SB: IV (Lidz, Webster, & Townes-Rosenwein, 1992). The BDI had weak predictive validity for children under 2 years old (Behl & Akers, 1996; Saylor, Boyce, Peagler, & Callahan, 2000), and identifying children with social emotional needs (Merrell & Mauk, 1993).

Mixed results were reported for the 8 studies that examined the effectiveness of the **BSID**. One study found little correlation between the BSID and BDI (Gerken, Eliason, & Arthur, 1994). While other studies found high correlations between the BSID-II Motor Scale and the Peabody Developmental Motor test (Provost et al., 2004), the BSID-II Mental and Motor Scales and the Play Based Assessment (Kelly-Vance et al., 1999), and the BSID-II and Merrill-Palmer (Magiati & Howlin, 2001). Crosby (1999) examined the effectiveness of the **MSEL** and found it a statistically valid and reliable tool in the assessment of Hispanic infants.

Six studies examined the effectiveness of the **SB** and results were mixed. These studies showed evidence for concurrent validity of the SB, but not predictive validity. Moderate to high correlations were found between the SB-IV, LIPS and Vineland, the SB and WPPSI, and the SB: IV and BDI (Atkinson, Beve, Dickens, & Blackwell, 1992; Gerken & Hodapp, 1992; Lidz, Webster, & Townes-Rosenwein, 1992). Grunau and colleagues (2000) found the SB: IV to be a poor predictor of 3-year old children's skills at ages 4 and 5. The authors noted the standard method of calculating the standard score on the SB: IV excludes subtests with a raw score of 0, which over estimates cognitive functioning in young children. Similarly, Saylor, Boyce, Peagler, and Callahan (2000) found the SB: IV only correctly identified 13% of children found by the BDI to be delayed.

Support for the effectiveness of the **WPPSI** was reported by 2 studies. These studies showed evidence for concurrent validity and construct validity. High correlations were found between the WPPSI-R and SB, ranging from .75 to .85 (Gerken & Hodapp, 1992). Ottem (2003) reported evidence for the construct validity of the WPPSI (Ottem, 2003).

*Effectiveness related to test characteristics.* An effective test performs the function for which it is intended. Conventional tests aim to identify individuals with delay or disability. The

extent to which the research reviewed in this synthesis investigated *effectiveness* was directly influenced by the presence or absence of the six essential test characteristics.

Children with disabilities were not included in the standardization samples in the tests which were the focus of the studies reviewed in this synthesis. The absence of a disability sample raises questions about the social effectiveness of the tests and research results. Social expectations about the functioning of a child with a disability are subject to shift as time goes by. Take for example the advent of new technologies (e.g., voice output devices, mobility aids, etc.) and medical breakthroughs that allow a child with a disability to function more independently than in the past. Today, children with disabilities have access to community environments that are inclusive and represent their natural everyday settings, as protected by legislative actions. When children with disabilities are excluded from the normative sample, social validity of the test is dubious at best because the developmental performance of children with disabilities may be over- or under- estimated when using norms developed solely on their typically developing peers. An unfortunate result could be invalid eligibility determination, which could lead to inappropriate early intervention service decisions. Effectiveness of test outcomes improves when tests have a normative sample which includes people with disabilities.

Procedural flexibility was another missing test characteristic in the tests that were used in the research studies we reviewed. Lack of procedural flexibility was problematic for supporting effectiveness of conventional tests, because a few of the concurrent validity studies had weak to moderate classification agreement and correlation coefficients (Gerken et al., 1994; Gerken & Hodapp, 1992; Saylor et al., 2000). Inflexible application of standardized procedures may result in unfair and ineffective decisions about a child's eligibility status, because young children with and without disabilities may not be interested in the testing session, easily distracted, or need

alternative procedures and accommodations to meet their unique developmental needs. Forcing a young child to participate in test activities that are not procedurally flexible remains a questionable practice that may be ineffective for many children.

A test used to determine eligibility for early intervention should cover multiple areas of a young child's development in order to effectively understand his/her unique developmental and educational needs. Coverage across multiple developmental areas makes it possible to determine if a child would benefit from specialized services which will address a variety of developmental domains (e.g., adaptive, cognitive, communication, fine motor, gross motor, and social). The presence of comprehensive coverage on a handful of tests had encouraging outcomes for some of the concurrent validity studies in terms of moderate to strong classification agreement between tests and correlation coefficients (Boyd et al., 1989; Guidubaldi & Perry, 1984; Hurt et al., 2001; Johnson et al., 1992; Kelly-Vance et al. 1999; Lidz et al., 1992; Magiati & Holin, 2001; McLean et al., 1987; Provost et al., 2004; Tingey et al., 1991). Conventional tests are better equipped to answer the eligibility question when they broadly comprise areas that are intended to identify delay or disability.

An eligibility test is more effective and useful when it can be used to inform the delivery of early intervention services. Treatment validity studies would help us better understand if conventional tests are able to effectively be used by test consumers to write goals for intervention, create/implement intervention content and curricula, and ultimately monitor whether children are making progress from their initial eligibility assessment or baseline performance data. When tests contain functional content, the items (i.e., content) could be used for programmatic purposes. Unfortunately, however, we did not find any treatment validity studies in our review of the literature. We located two utility studies that showed how useful

results could be obtained for children with mild/moderate disabilities and diverse cultural/linguistic backgrounds when tests had functional content (Bailey et al., 1987; Crosby, 1999).

In addition to functional test content, the efficacy of conventional tests improves with the existence of a collection of multiple and varied test items. *Item density* enhances the ability of the test to effectively determine if a child is eligible for early intervention through the richness of skills that are displayed on the test, especially in the early years of development. Scoring the child's performance should also be taken into account when determining eligibility and test efficacy. *Graduated scoring* builds on the issue of item density by offering the ability to describe the child's test performance in a more complex manner. For example, a three-point rating scale could provide more information than a dichotomous scoring feature, because there can be opportunities for varying levels of skill acquisition that can be better represented with a graduated scoring option. The number of items and how those items are scored are important considerations when making decisions because this also impacts the ability of a test to translate into information that is useful beyond answering the simple eligibility question (i.e., "is this child eligible for early intervention?"). Conventional tests were missing the qualities of *item density* and *graduated scoring*. The absence of these two test characteristics had a negative impact on studies because they were unable to effectively identify very young children under the age of 2 for early intervention or predict performance (Behl & Akers, 1996; Grunau et al., 2000; Merrell & Mauk, 1993; Saylor et al., 2000). Table 5 shows the results of the 29 studies conducted on the accuracy and effectiveness of the conventional tests used to assess young children.

<insert Table 5 here>

## Conclusion

There are possible limitations of the current synthesis. We examined nine conventional tests, which included their various updated editions. A number of conventional tests are commercially available, however we chose these tests because they appeared most often in the professional literature and by accounts from practitioners. The body of literature contained other publications on conventional testing. We only included research studies that involved young children who were at risk or had a disability in the sample.

Test characteristics were identified to provide a framework for measuring children's strengths and needs in a reliable, representative, comprehensive, accurate, and useful fashion. The test characteristics indicate best practice recommendations for meeting legal and procedural requirements, and that are developmentally-appropriate for young children. Bagnato and Neisworth (1991) proposed a definition for early childhood assessment that is noted in DEC Recommended Practices (2005): *Early childhood assessment is a flexible, collaborative decision making process in which teams of parents and professionals repeatedly revise their judgments and reach consensus about the changing developmental educational, medical, and mental health service needs for young children and their families.*

When assessing young children, the strengths and needs of the child and family must be considered and are likely to influence any assumption about the amount of progress that may or may not take place. It is important that the test measures what it is intended to measure. The consistency of assessing children is especially significant at transition times as the criteria of eligibility changes but continuity of services is of high importance for the child and the families involved. Two areas were addressed in this practice-based research synthesis which included

conventional test characteristics, and research support for the use of conventional testing of young children with disabilities.

### *Implications for Practice*

Conventional tests are evolving in order to attend to the early intervention eligibility standards of IDEA (2004). The BDI-2, BSID-III, SB5, and WPPSI-III are revised editions developed in order to help meet the IDEA standards. For example, the test manuals direct administrators to incorporate parent participation in the evaluation process. The BDI-2 and BSID-III assess domains that are aligned with IDEA regulations. Further, the test manuals emphasize the importance of cultural sensitivity and allow for some standardization modifications when assessing children with disabilities. While these updates are promising, we have additional suggestions for improving practices.

First, conventional tests should contain a representative disability within the norming sample. Unfortunately, there were only two tests (22%) in our review that included young children with disabilities in the standardization sample. One of the purposes of assessment is to determine eligibility for early intervention. It is critical that the test was normed on the population of children similar to those for whom it is to be used. Furthermore, it is necessary that children with disabilities were included in the standardization sample and norm referencing of the measure.

Second, conventional tests should allow for procedural flexibility. We found that there were no tests in our review that had procedural flexibility. It is important to have flexibility when assessing children in order to enable parents to participate, to allow for professional collaboration, to administer the assessment in a natural environment, to be culturally sensitive, to have opportunities to adapt materials when necessary, and to use informed clinical opinion to

estimate if the adaptations measure the same skill as it is intended. Procedural flexibility allows for flexibility in administration, modifications to assessment for use with children who have specific impairments and delays. This includes flexibility when performing the assessment, scoring and other accommodations that may be used to demonstrate a child's knowledge and function of skills. Accommodations made in order for a child to accomplish a skill, and how the accommodation created by the assessor's clinical judgment affects reliability are important considerations. When interacting with younger children it sometimes is important to follow the child's lead, therefore flexibility in the assessor's approach is necessary.

Third, conventional tests should cover multiple developmental domains. Less than half of the tests in our review included a comprehensive arrangement of domains. Additionally, there were only a couple of tests that included a variety of score types and none that met our standards for item density.

The fourth and final suggestion is that assessments used for eligibility determination contain functional items. There were only two conventional tests (22%) that included some functional response items in our review. We stated earlier in this synthesis that an advantage of conventional tests is that they answer the question, "Is this child eligible for services?" However, a disadvantage of many conventional tests is that they are not able to go beyond that question to identify goals and intervention content for the child, because their items are not authentic skills that the child needs when they encounter everyday, real life experiences. Very few conventional tests are appropriate for programmatic purposes (i.e., goal development, intervention, and program evaluation). By only being able to answer the yes/no question about eligibility, conventional tests are limited.

*Implications for Research*

Assessments should be reliable in order to determine standard deviation and/or age equivalent to establish eligibility. Reliability of the assessment tool is the consistency of an assessment tool (being free of error), containing consistency across test items, and use of cut off score in order for the tool to precisely or accurately measure the true attributes of a child. Validity addresses to what extent the assessment tool measures what it was designed to measure, it relates significantly to similar measures and discriminates among special populations of children. This may include content, construct, processes, relation to other variables, sensitivity/specificity, internal structure and consequence. Interestingly, in our synthesis we discovered that there was more research on validity (90%) than reliability (34%). Both are important psychometric properties of a test. Part of the decision to determine what tool to use for assessing a child should take validity into consideration. The scores (i.e. standard scores, age-equivalent scores) the assessment produces is also an important factor when determining eligibility for early intervention services.

Research is needed to examine the: (1) extent conventional tests are used for Part B and Part C eligibility decisions, (2) ways in which conventional tests are used, and (3) degree to which conventional tests facilitate eligibility, placement, and early programming decisions. Conventional intellectual measures are less reliable with infants and preschoolers and no more functional in providing data for interventions and progress monitoring than with older students. Because a conventional test by itself is not adequate to determine all the functions needed, research efforts could identify which combination of tools best determines eligibility for young children and their families.

In addition, social and treatment validity studies are needed to compare the effectiveness of conventional tests and how different types of assessments help to develop aspects of early intervention including: parent participation, cultural considerations (bias free), natural environment, providing information helpful in creating an IFSP/IEP, providing a beginning curriculum for program providers, and providing input for program evaluation.

Assessment practices should facilitate decision making in the eligibility gateway to IDEA early intervention services. Given a finite amount of resources, it is difficult for the consumer to locate the best possible assessment that will be useful in determining eligibility that leads to the development of meaningful goals, effective intervention, and ultimately program evaluation. This synthesis identified test characteristics and relevant research characteristics that apply to early intervention, which will assist in the process of making informed decisions about eligibility assessment.

## References

\* Indicates *studies* used in the synthesis

\*\* Indicates *measures* used in the synthesis

American Educational Research Association (1999). *Standards for educational and psychological testing*. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Washington, DC: AERA.

\*Atkinson, L., Bevc, I., Dickens, S., & Blackwell, J. (1992). Concurrent validities of the Stanford-Binet (Fourth Edition), Leiter, and Vineland with developmentally delayed children. *Journal of School Psychology, 30*(2), 165-173.

Bagnato, S. J., & Neisworth, J. T. (1991). *Assessment for early intervention: Best practices for professionals*. New York: Guilford.

Bagnato, S. J., & Neisworth, J. T. (1994). A national study of the social and treatment "invalidity" of intelligence testing for early intervention. *School Psychology Quarterly, 9*(2), 81-102.

Bagnato, S. J., & Neisworth J.T. (2005). DEC Recommended Practices: Assessment. In S. Sandall, M.L. Hemmeter, B.J. Smith, & M. McLean (Eds.), *DEC Recommended Practices* (pp.45-50). Longmont, CO: Sopris West.

\*Bailey, D. B., Vandiviere, P., Dellinger, J., & Munn, D. (1987). The Battelle Developmental Inventory: Teacher perceptions and implementation data. *Journal of Psychoeducational Assessment, 5*(3), 217-226.

Bailey, D.B., & Wolery, M. (1989). *Assessing infants and preschoolers with handicaps*. Columbus, OH: Merrill Pub. Co.

- \*\*Bayley, N. (1993). *Scales of infant development second edition manual*. San Antonio, TX: The Psychological Corporation.
- \*\*Bayley, N. (2006). *Bayley scales of infant and toddler development third edition manual*. San Antonio, TX: Harcourt.
- \*Behl, D. D., & Akers, J. F. (1996). The use of the Battelle Developmental Inventory in the prediction of later development. *Diagnostic*, 21(4), 1-16.
- \*Boyd, R. D., Welge, P., Sexton, D., & Miller, J., H. (1989). Concurrent validity of the Battelle Developmental Inventory: Relationship with the Bayley Scales in young children with known or suspected disabilities. *Journal of Early Intervention*, 13(1), 14-23.
- Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment*, 4, 313-326.
- Bradley-Johnson, S. (2001). Cognitive assessment for the youngest children: A critical review of tests. *Journal of Psychoeducational Assessment*, 19, 19-44.
- \*Cook, M. J., Holder-Brown, L., Johnson, L. J., & Kilgo, J. L. (1989). An examination of the stability of the Bayley Scales of Infant Development with high-risk infants. *Journal of Early Intervention*, 13(1), 45-49.
- \*Crosby, F. X. (1999). A comparative study of the Mullen Scales of Early Learning with Hispanic infants. Dissertation Abstracts International: Section B: The Sciences and Engineering. Vol 60(4-B), Miami.
- Danaher, J. (2005). *Eligibility policies and practices for young children under Part B of IDEA* (NECTAC Notes No.15). Chapel Hill: The University of North Carolina, FPG Child Development Institute, National Early Childhood Technical Assistance Center.

- \*Dezoete, J. A., MacArthur, B. A., & Tuck, B. (2003). Prediction of Bayley and Stanford-Binet scores with a group of very low birthweight children. *Child: Care, Health and Development, 29*(5), 367-372.
- Dunst, C.J., Trivette, C.M., & Cutspe, P.A. (2002). An evidence-based approach to documenting the characteristics and consequences of early intervention practices. *Centerscope, 1*(2), 1-6. Available at <http://www.evidencebasepractices.org/centerscope/centerscopevol1no2.pdf>.
- Flanagan, D., & Alfonso, V. (1995). A critical review of the technical characteristics of new and recently revised intelligence tests for preschool children. *Journal of Psychoeducational Assessment, 13*, 66-90.
- \*Gerken, K. C., Eliason, M. J., & Arthur, C. R. (1994). The assessment of at-risk infants and toddlers with the Bayley Mental Scale and the Battelle Developmental Inventory: Beyond the data. *Psychology in the Schools, 31*(3), 181-187.
- \*Gerken, K. C., & Hodapp, A. F. (1992). Assessment of preschoolers at-risk with the WPPSI--R and the Stanford-Binet L-M. *Psychology Report, 71*(2), 659-664.
- Goh, D. S., Teslow, C. J., & Fuller, G. B. (1981). The practice of psychological assessment among school psychologists. *Professional Psychology, 12*(6), 696-704.
- \*Grunau, R. E., Whitfield, M. F., & Petrie, J. (2000). Predicting IQ of biologically "at risk" children from age 3 to school entry: sensitivity and specificity of the Stanford-Binet Intelligence Scale IV. *Journal of Developmental & Behavioral Pediatrics, 21*(6), 401-407.
- \*Guidubaldi, J., & Perry, J. D. (1984). Concurrent and Predictive Validity of the Battelle Development Inventory at the First Grade Level. *Educational and Psychological Measurement, 44*(4), 977-985.

- \*Hatton, D. D., Bailey, D. B., Burchinal, M. R., & Ferrell, K. A. (1997). Developmental growth curves of preschool children with vision impairments. *Child Development, 68*(5), 788.
- \*Hurt, H., Malmud, E., Betancourt, L. M., Brodsky, N. L., & Giannetta, J. M. (2001). A prospective comparison of developmental outcome of children with in utero cocaine exposure and controls using the Battelle Developmental Inventory. *Journal of Developmental & Behavioral Pediatrics, 22*(1), 27-34.
- Hutton, J. B., Dubes, R., & Muir, S. (1992). Assessment practices of school psychologists: Ten years later. *School Psychology Review, 21*(2), 271.
- \*Johnson, L. J., Cook, M. J., & Kullman, A. J. (1992). An examination of the concurrent validity of the Battelle Developmental Inventory as compared with the Vineland Adaptive Scales and the Bayley Scales of Infant Development. *Journal of Early Intervention, 16*(4), 353-359.
- \*Kelly-Vance, L., Needelman, H., Troia, K., & Ryalls, B.O. (1999). Early childhood assessment: A comparison of the Bayley Scales of Infant Development and play-based assessment in two-year old at-risk children. *Developmental Disabilities Bulletin, 27*(1), 1-15.
- The Individuals with Disabilities Education Act, 20 U.S.C.§1414(1)-(3), 1412(a)(6)(B)(2004)
- \*Lidz, C. S., Webster, I., & Townes-Rosenwein, L. (1992). Concurrent validity of the cognitive domain of the Battelle Developmental Inventory in relation to the Stanford-Binet Intelligence Test, Fourth Edition for urban African-American low SES preschool children. ERIC ED350344.
- \*Magiati, I. & Howlin, P. (2001). Monitoring the progress of preschool children with autism enrolled in early intervention programs. *Autism, 5*(4), 399-406.

- Mardell-Czudnowski, C. (1996). A survey of assessment professionals in the US (Testing children with special needs). *School Psychology International, 17*, 189-209.
- \*Mayes, S. D. (1997). Potential Scoring Problems Using the Bayley Scales of Infant Development-II Mental Scale. *Journal of Early Intervention, 21*(1), 36-44.
- \*McConachie, H., Le Couteur, A., & Honey, E. (2005). Can a diagnosis of asperger syndrome be made in very young children with suspected autism spectrum disorder? *Journal of Autism and Developmental Disorders, 35*, 167-176.
- McLean, M., Bailey, D.B., & Wolery, M. (2004). *Assessing infants and preschoolers with special needs (3rd ed.)*. Upper Saddle River, NJ: Pearson.
- \*McLean, M., McCormick, K., Bruder, M. B., & Burdug, N. B. (1987). An investigation of the validity and reliability of the Battelle Developmental Inventory with a population of children younger than 30 months with identified handicapping conditions. *Journal of the Division for Early Childhood, 11*(3), 238-246.
- \*Merrell, K. W., & Mauk, G. W. (1993). Predictive validity of the Battelle Developmental Inventory as a measure of social-behavioral development for young children with disabilities. *Diagnostic, 18*(3), 187-198.
- \*Mott, S. E. (1987). Speech and language disordered children. *Psychology in the Schools, 24*, 215-220.
- \*\*Mullen, E. M. (1995). *Mullen scales of early learning*. Circle Pines, MN: American Guidance Service Inc.
- NASP (2005). *NASP position statement on early children assessment*. Bethesda, MD: Association of School Psychologists.

- \*\*Newborg, J., Stock, J. R., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984). *Battelle developmental inventory*. Allen, TX: DLM/Teaching Resources.
- \*\*Newborg, J. (2005). *Battelle developmental inventory second edition examiner's manual*. Itasca, IL: Riverside.
- \*Ottem, E. (2003). Confirmatory factor analysis of the WPPSI, WPPSI-R, and the WISC-R: Evaluation of a model based on knowledge-dependent and processing-dependent subtests. *Journal of Psychoeducational Assessment, 21*(1), 3-15.
- Pretti-Frontczak, K., Kowalski, K., & Douglas-Brown, R. (2002). Preschool teachers' use of assessments and curricula: A statewide examination. *Exceptional Children, 69*(1), 109-
- \*Provost, B., Heimerl, S., McClain, C., Kim, N. H., Lopez, B. R., & Kodituwakku, P. (2004). Concurrent validity of the Bayley Scales of Infant Development II Motor Scale and the Peabody Developmental Motor Scales-2 in children with developmental delays. *Pediatric Physical Therapy, 16*(3), 149-156.
- \*\*Roid, G. H. (2006). *Stanford-Binet Intelligence Scales: Fifth edition*. Chicago: Riverside.
- \*Ross, G. (1985). Use of the Bayley Scales to characterize abilities of premature infants. *Child Development, 56*(4), 835-842.
- Sattler, J. M. (2001). *Assessment of children: Cognitive applications (4th ed.)*. San Diego: Jerome M. Sattler, Publisher, Inc.
- \*Saylor, C. F., Boyce, G. C., Peagler, S. M., & Callahan, S. A. (2000). Brief report: cautions against using the Stanford-Binet-IV to classify high-risk preschoolers. *Journal of Pediatric Psychology, 25*(3), 179-183.

- \*Sexton, D., McLean, M., Boyd, R., Thompson, B., & McCormick, K. (1988). Criterion-related validity of a new standardized developmental measure for use with infants who are handicapped. *Measurement and Evaluation in Counseling and Development, 21*, 16-24.
- Shackelford, J. (2006). *State and jurisdictional eligibility definitions for infants and toddlers with disabilities under IDEA* (NECTAC Notes No. 21). Chapel Hill: The University of North Carolina, FPG Child Development Institute, National Early Childhood Technical Assistance Center.
- \*Snyder, P., Lawson, S., Thompson, B., Stricklin, S., & Sexton, D. (1993). Evaluating the psychometric integrity of instruments used in early intervention research: The Battelle Developmental Inventory. *Topics in Early Childhood Special Education, 13*(2), 216-232.
- \*\*Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scales: Fourth edition*. Chicago: Riverside.
- \*Tingey, C., Mortensen, L., Matheson, P., & Doret, W. (1991). Developmental attainment of infants and young children with Down syndrome. *International Journal of Disability, Development and Education, 38*(1), 15-26.
- \*\*Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence- Revised*. San Antonio, TX: The Psychological Corporation.
- \*\*Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence-Third Edition*. San Antonio, TX: The Psychological Corporation.
- Wilson, R.S. (1983). The Louisville twin study: Developmental synchronics in behavior. *Child Development, 54*(2), 298-316.
- Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review, 25*(1), 9-23. *Childhood Special Education, 24*(2), 110-120.

U.S. Department of Education Office of Elementary & Secondary Education. (2002). *No Child Left Behind. A Desktop Reference*, Washington, DC: U.S. Government Printing Office.

Table 1

*Conventional Test Characteristics*

Conventional Test ( <i>N</i> = 9)	Standardization Includes Children with Disabilities	Procedural Flexibility	Comprehensive Coverage	Functional Content	Item Density	Graduated Scoring
BDI-1	No	No	<b>Yes</b>	No	No	<b>Yes</b>
BDI-2	<b>Yes</b>	No	<b>Yes</b>	<b>Yes</b>	No	<b>Yes</b>
BSID-II	No	No	No	No	No	No
BSID-III	<b>Yes</b>	No	<b>Yes</b>	No	No	No
MSEL	No	No	<b>Yes</b>	<b>Yes</b>	No	No
SB: IV	No	No	No	No	No	No
SB5	No	No	No	No	No	No
WPPSI-R	No	No	No	No	No	No
WPPSI-III	No	No	No	No	No	No

*Note.* BDI = Battelle Developmental Inventory; BSID = Bayley Scales of Infant Development; MSEL = Mullen Scales of Early Learning, AGS Edition; SB = Stanford Binet Intelligence Scales; WPPSI = Wechsler Preschool & Primary Scales of Intelligence.

Table 2

*Five Conventional Tests with Multiple Versions*

BDI-1	BDI-2
5 domains: <ul style="list-style-type: none"> <li>• Adaptive</li> <li>• Cognitive</li> <li>• Communication</li> <li>• Motor</li> <li>• Personal-Social</li> </ul>	5 domains: <ul style="list-style-type: none"> <li>• Adaptive</li> <li>• Cognitive</li> <li>• Communication</li> <li>• Motor</li> <li>• Personal-Social</li> </ul>
22 sub-domains	13 sub-domains
341 items	450 items
Age range: birth to 7.11 years	Age range: birth to 7.11 years
Norms: $N = 800$	Norms: $N = 2,500$
Children with disabilities: The manual does <b>NOT</b> report children with disabilities included in the standardization sample	Children with disabilities: YES Autism ( $n = 44$ ) Cognitive ( $n = 42$ ) Developmental delay ( $n = 58$ ) Motor ( $n = 40$ ) Premature birth ( $n = 45$ ) Speech and Language ( $n = 72$ )
Norm intervals: 6 mo (birth-23 months) 1 year (24 months and up)	Norm intervals: 1 month (birth to 23 month) 3 month (24 month and up)
Accommodations: YES	Accommodations: YES
Newborg, J., Stock, J.R., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984). <i>Battelle developmental inventory</i> . Allen, TX: DLM/Teaching Resources.	Newborg, J. (2005). <i>Battelle developmental inventory second edition examiner's manual</i> . Itasca, IL: Riverside.

BSID-II	BSID-III
<p>3 domains</p> <ul style="list-style-type: none"> <li>• Mental Scale</li> <li>• Psychomotor Scale</li> <li>• Behavior Rating Scales</li> </ul> <p>A separate Motor Scale Kit is available</p> <p>4 sub-domains</p> <p>319 items</p> <p>Age range: 1 to 42 months</p> <p>Norms: <math>N = 1,700</math></p> <p>Children with disabilities included: NO</p> <p>Norm intervals: 1 month for 36 months 3 months to 42 months</p> <p>Accommodations: None</p> <p>Bayley, N. (1993). <i>Scales of infant development second edition manual</i>. San Antonio, TX: The Psychological Corporation.</p>	<p>5 domains</p> <ul style="list-style-type: none"> <li>• Adaptive Behavior Scales</li> <li>• Cognitive</li> <li>• Language</li> <li>• Motor</li> <li>• Social-Emotional</li> </ul> <p>14 sub-domains</p> <p>567 items</p> <p>Age range: 16 days to 43 months 15 days</p> <p>Norms: <math>N = 1,700</math> Social-emotional scale: <math>N = 456</math> Adaptive Behavior scale: <math>N = 1,350</math></p> <p>Children with disabilities included: YES</p> <p>Norm intervals: 1month up to 6 months 2.5 or less to 13 months 15 days 3.5 or less to 28 months 16 days to 38 months 30 days 3.5 39 months to 42 months 15 days</p> <p>Accommodations: Not recommended. If clinical judgment is involved.</p> <p>Bayley, N. (2006). <i>Bayley scales of infant and toddler development – Third edition</i>. San Antonio, TX: Harcourt assessment, Inc.</p>

---

**MSEL**

---

2 domains

- Cognitive
- Gross motor

4 cognitive sub-domains: visual reception, fine motor, receptive language, and expressive language

159 items

Age range: Birth to 68 months

Norms: N=1,849

Children with disabilities included: No

Norm intervals in months:

1-2, 3-4, 5-6, 7-8, 9-10, 11-12, 13-14, 15-17, 18-20, 21-26, 27-32, 33-38, 39-44, 45-50, 51-56, 57-62 , and 63-68

Accommodations: NO.

Mullen, M.E. (1995). *Mullen Scales of Early Learning, AGS Edition*. Circle Pines, MN: American Guidance Service.

SB: IV	SB5
<p>A tiered factor system is used, which includes Tier 1: general intelligence, Tier 2: crystallized abilities, fluid analytic abilities, and short term memory, Tier 3: verbal reasoning, quantitative reasoning, and abstract/visual reasoning.</p>	<p>2 domains: verbal and nonverbal IQ</p>
	<p>5 sub-domains: fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing, and working memory.</p>
<p>Number of items not reported</p>	<p>293 items</p>
<p>Age range: 2 years through adult</p>	<p>Age range: 2 years through 89.11 years</p>
<p>Norms: <math>N=5,013</math> on ages 2-23</p>	<p>Norms: <math>N=4,800</math> <math>N=1,400</math> ages 2-5 years</p>
<p>Children with disabilities included: No</p>	<p>Children with disabilities included: Five percent of the sample included students in special education programs who were in regular education programs for 50% of the day.</p>
<p>Norm intervals: 17 total age groups. 2 to 17 in increments of 1 year, 18-23 as one age group.</p>	<p>Norm intervals: 30 age groups were defined. Children in the early childhood groups (ages 2 to 4 years) were divided into half year groupings each with <math>N=200</math> subjects in each.</p>
<p>Accommodations: NO. Unless standard procedures are followed, the results lose their significance.</p>	<p>Accommodations: NO. Accommodations may change the standardized nature of test interpretation. Results should be interpreted with caution if a disability is known. Cautions for specific disabilities are provided.</p>
<p>Thorndike, R.L., Hagen, E.P., &amp; Sattler, J.M. (1986). <i>The Stanford-Binet Intelligence Scale: Fourth edition</i>. Chicago, IL: Riverside.</p>	<p>Roid, G. H. (2006). <i>Stanford-Binet Intelligence Scale for Early Childhoods: Fifth edition</i>. Chicago: Riverside.</p>

WPPSI-R	WPPSI-III
3 composite scores	Ages 2:6-3:11 4 possible composite scores Ages 4:0-7:3 5 possible composite scores
12 subtests	Ages 2:6-3:11 4 core subtests + 1 supplemental Ages 4:0-7:3 7 core subtests + 5 supplemental + 2 optional
Age range: 3 years to 7 years 3 months	Age range: 2 years 6 months to 7 years 3 months
Norms: $N = 1,200$ children	Norms: $N = 1,700$ children
Children with disabilities included: NO	Children with disabilities included: NO Special group studies included in reliability and validity analyses
Norm intervals: 16 three-month intervals from 2.11 to 6.11 1 four-month interval from 6.11 to 7.3	Norm intervals: 2.6 to 2.11 3.0 to 3.5 3.6 to 3.11 4.0 to 4.5 4.6 to 4.11 5.0 to 5.5 5.6 to 5.11 6.0 to 6.11 7.0 to 7.3
Accommodations: NO Modifications may be made but must be documented.	Accommodations: NO Modifications may be made but must be documented.
Wechsler, D. (1989). <i>Wechsler Preschool and Primary Scale of Intelligence-Revised</i> . San Antonio, TX: The Psychological Corporation.	Wechsler, D. (2002). <i>Wechsler Preschool and Primary Scale of Intelligence-Third edition</i> . San Antonio, TX: The Psychological Corporation.

Table 3

*Research Studies with Participant Demographic Information (N = 29)*

<b>Study Reference</b>	<b>Sample Size</b>	<b>Mean Age in Months</b>	<b>Age Range in Months</b>	<b>Child Characteristics</b>
Atkinson, Bevc, Dickens, & Blackwell, 1992	24	81.60	48-132	Developmental delay
Bailey, Vandiviere, Dellinger, & Munn, 1987	247	47	11-92	Mental retardation, at-risk, developmental delay, & speech and language
Behl & Akers, 1996	239	NR	6-52	Developmental delay or mild/moderately mentally retarded, low incidence
Boyd, Welge, Sexton, & Miller, 1989	30	15.07	0-30	Known or suspected disability
Cook, Holder-Brown, Johnson, Kilgo 1989	80	NR	6-12	Hospitalization in the NICU for 2 wks to 6 months- developmentally at-risk due to medical condition
Crosby, 1999	65 Hispanic 68 non-Hispanic	NR	0-36	Spanish-dominant language homes
Dezoete, MacArthur, & Tuck 2003	334	18 - 18.50 48 - 48.40	18 & 48	Low birth weight
Gerken, Eliason, & Arthur, 1994	34	13.80	3-30	At risk: teen parents
Gerken & Hodapp, 1992	16	NR	36-60	Mild to moderate mental disability and physical disability
Grunau, Whitfield, & Petrie, 2000	236	NR	36-54	Biologically at risk children without major sensory or motor impairments
Guidubaldi & Perry, 1984	50	NR	NR	At risk: low SES
Hatton, Bailey, Burchinal, & Ferrell, 1997	186	38	12-73	Visual disability and MR/DD
Hurt, Malmud, Betancourt, Brodsky, & Giannetta, 2001	133	37 & 61	36-47 & 60-71	At risk: children exposed to cocaine in utero
Johnson, Cook, Kullman, 1992	67	18.63	2-60	Motor
Kelly-Vance, Needelman, Troia, & Ryalls, 1999	38	24	23-27	Low birth weight and required medical intervention during first month of life

Lidz, Webster, & Townes-Rosenwein, 1992	32	50.90	38-62	At risk: mild to moderate learning and/or behavioral concerns
Magiati & Howlin, 2001	24	39.60	27-58	Autism
Mayes, 1997	32	25	5 to 77	Brain injury, autism, cerebral palsy, metabolic disorders, FAS, hydrocephalus, spina bifida, visual impairments
McConachie, Couteur, & Honey, 2005	104	NR	24-48	Autism spectrum
McLean, McCormick, Bruder, & Burdg, 1987	40	16.53	2-28	Sensory, developmental delay, physical or multiple disabilities
Merrell & Mauk, 1993	121	60.54	60-96	Developmental delay or mild/moderately mentally retarded, low incidence
Mott, 1987	20	49.50	35-60	Speech and language
Ottem, 2003	198	5.60	NR	Language impaired
Provost, Heimerl, McClain, Kim, Lopez, & Kodituwakku, 2004	110	25.30	3-41	79% developmental delays in at least 2 areas (cognitive, speech, motor); 9% speech delay; 7% motor delay; 18% autism/PDD
Ross, 1985	92 infants 46 premature 46 full term	12	NR	Low birth weight
Saylor, Boyce, Peagler, & Callahan, 2000	92	NR	36-60	At risk: all children had experienced intraventricular hemorrhage and/or low birth weight less than 1000 grams
Sexton, McLean, Boyd, Thompson, & McCormick, 1988	70	16.24	0-30	Developmental delay
Snyder, Lawson, Thompson, Stricklin, & Sexton, 1993	78	6 to 11 11.00 12 to 17 14.80 18 to 23 20.80 24 to 35 29.30 36 to 47 38.60 48 to 59 52.80 60 to 71 63.30 72 to 83 73.80	0-95	Severe disabilities
Tingey, Mortensen, Matheson, & Doret, 1991	198	NR	0-70	Down syndrome

---

*Note.* NR = not reported.

Table 4

*Research Study Characteristics*

Studies (N = 29)	Test	Accuracy (Reliability)			Effectiveness (Validity)			
		Inter-item <sup>a</sup> Inter-rater <sup>b</sup>	Test-retest	Sensitivity <sup>a</sup> Specificity <sup>b</sup>	Concurrent	Construct <sup>a</sup> Criterion <sup>b</sup>	Predictive	Utility
Atkinson, Bevc, Dickens, & Blackwell, 1992	SB: IV				X			
Bailey, Vandiviere, Dellinger, & Munn, 1987	BDI							X
Behl & Akers, 1996	BDI						X	
Boyd, Welge, Sexton, & Miller, 1989	BDI & BSID	X <sup>b</sup>	X		X			
Cook, Holder-Brown, Johnson, Kilgo 1989	BSID		X					
Crosby, 1999	MSEL							X
Dezoete, MacArthur, & Tuck, 2003	BSID-II & SB: IV						X	
Gerken, Eliason, & Arthur, 1994	BDI & BSID				X			
Gerken & Hodapp, 1992	SB & WPPSI-R				X			
Grunau, Whitfield, & Petrie, 2000	SB: IV & WPPSI-R			X			X	
Guidubaldi & Perry, 1984	BDI				X		X	
Hatton, Bailey, Burchinal, & Ferrell, 1997	BDI	X <sup>b</sup>						
Hurt, Malmud, Betancourt, Brodsky, & Giannetta, 2001	BDI				X			
Johnson, Cook, & Kullman, 1992	BDI & BSID				X			
Kelly-Vance, Needelman, Troia, & Ryalls, 1999	BSID-II			X	X			
Lidz, Webster, & Townes-Rosenwein, 1992	BDI & SB: IV				X			
Magiati & Howlin, 2001	BSID-II				X			
Mayes, 1997	BSID-II			X <sup>a</sup> X <sup>b</sup>				

McConachie, Couteur, & Honey, 2005	MSEL		X <sup>a</sup> X <sup>b</sup>		
McLean, McCormick, Bruder, & Burdg, 1987	BDI & BSID	X <sup>a</sup> X <sup>b</sup>		X	
Merrell & Mauk, 1993	BDI				X
Mott, 1987	BDI			X <sup>a</sup>	
Ottem, 2003	WPPSI-R			X <sup>a</sup>	
Provost, Heimerl, McClain, Kim, Lopez, & Kodituwakku, 2004	BSID-II			X	
Ross, 1985	BSID				X
Saylor, Boyce, Peagler, & Callahan, 2000	BDI & SB: IV			X	X
Sexton, McLean, Boyd, Thompson, & McCormick, 1988	BDI & BSID	X <sup>b</sup>		X <sup>b</sup>	
Snyder, Lawson, Thompson, Stricklin, & Sexton, 1993	BDI	X <sup>a</sup> X <sup>b</sup>		X <sup>a</sup>	
Tingey, Mortensen, Matheson, & Doret, 1991	BDI			X	

---

*Note.* BDI= Battelle Developmental Inventory; BSID= Bayley Scales of Infant Development; MSEL= = Mullen Scales of Early Learning, AGS Edition; SB= Stanford Binet–Fourth Edition; and WPPSI= Wechsler Preschool and Primary Scale of Intelligence.

Table 5

*Research Study Results (N = 29)*

Study Reference	Results
Atkinson, Bevc, Dickens, & Blackwell, 1992	<ol style="list-style-type: none"> <li>1. Significant correlation between the SB and LIPS, however intraindividual discrepancies were large and significant.</li> <li>2. Moderately significant correlation between the SB, LIPS and VABS scores, but mean score differences and intraindividual differences were large and significant.</li> <li>3. Vineland tends to provide a higher global score than either of the other 2 measures.</li> </ol>
Bailey, Vandiviere, Dellinger, & Munn, 1987	Teachers rated the BDI significantly more useful for children with mild/moderately handicapped, and not as useful for children with severe disabilities.
Behl & Akers, 1996	BDI consistently predicted WJR-ACH scores when the child was 24 months or older, however weak results for children 18 months or younger.
Boyd, Welge, Sexton, & Miller, 1989	Overlapping domains between the BDI and BSID.
Cook, Holder-Brown, Johnson, & Kilgo, 1989	Significant reliability coefficients between 6 month Mental Scale and Motor Scale scores and 12 month Mental Scale and Motor Scale scores. Average raw scores of at-risk infants were sig. lower on both Mental and Motor Scale, and half had scores more than 2 standard deviations below the mean standard score on both scales.
Crosby, 1999	Results suggest no significance for the ethnic variable and no significance for interaction of age and ethnicity. An affect was found for age only, where the younger the child in this <i>study</i> , the better they performed in most of the cognitive scales. This study reinforces the MSEL as a statistically valid and reliable tool in the assessment of Hispanic infants.
Dezoete, MacArthur, & Tuck, 2003	Children with VLBW had sig. higher scores on BSID-II at 18 months and S-B at 4 yrs. than children with ELBW. Longer term children had sig. higher scores on BSID-II at 18 mos. And S-B at 4 yrs. than shorter term children. Females' scores sig. higher on both measures. Systematic positive relationship found between SES and cognitive scores (higher SES had higher cognitive scores).
Gerken, Eliason, & Arthur, 1994	Study found little correlation between the BSID and BDI - .03.
Gerken & Hodapp, 1992	Correlations between WPPSI-R and SB-L-M ranged from .75 to .85.
Grunau, Whitfield, & Petrie, 2000	Using standard scoring, children with full scale IQs of less than 84 on the WPPSI at ages 4-5 were poorly identified (sensitivity of 54%) from the composite SB: IV score at age 3. Sensitivity improved to 78% by including as a predictor the number of subtests the child was actually able to perform at age 3. Measures from the Home Screening Questionnaire and ratings of mother-child interaction further improved sensitivity to 83%. The standard method of calculating the standard score on the SB: IV excludes subtests with a raw score of 0, which over estimates cognitive functioning in young biologically high risk children. Accuracy of early identification was improved by considering the number of subtests the child did not perform at age 3.
Guidubaldi & Perry, 1984	Data were interpreted to support the BDI as a valid multifactored assessment for use with at risk young children.
Hatton, Bailey, Burchinal, & Ferrell, 1997	The visual function and co-occurring disabilities had additive not multiplicative effects on the BDI Overall Developmental Age scores during early childhood.
Hurt, Malmud, Betancourt, Brodsky, & Giannetta, 2001	There were significant correlations between the Total HOME score and Total BDI at .23 for 3-Year and .52 for 5-Year.
Johnson, Cook, & Kullman, 1992	This was an extension of the McLean et al. (1987) study. Correlations were consistently lower than those found in the McLean study.
Kelly-Vance, Needelman, Troia, & Ryalls, 1999	Children had sig. higher scores on Play-Based Assessment (PBA) compared to BSID-II MDI score. 82% of children would not have been identified for EI services based on results of both assessments. 10% would have qualified using both techniques. 8% would have qualified from BSID-II MDI and not from PBA score. Sig. correlation found between the PBA, MDI, and PDI (motor) scores.

- Lidz, Webster, & Townes-Rosenwein, 1992 Evidence of a moderate positive relationship between cognitive domains of the BDI and SB. Authors caution that the BDI may not be a good measure for low functioning, mildly disabled preschoolers in all areas except possibly the cognitive domain.
- Magiati & Howlin, 2001 Standard scores on BSID-II and Merrill-Palmer highly correlated (.82). Standard scores on BSID-II sig. lower than scores on Merrill-Palmer. BSID-II and Vineland IQ equivalent scores similar (only sig. difference on Daily Living Skills) . Merrill-Palmer standard scores sig. higher than Vineland scores. The change in IQ scores was much greater for the children first tested on the BSID-II and then on the Merrill-Palmer, than those tested on the MP on both occasions.
- Mayes, 1997 When starting at item set closest to child's CA, the majority of BSID-II DA's were not lower than the BSID DA's. For the majority of children, the BSID-II MDI's obtained when began at the item set closest to the child's CA were above the expected range predicted from the BSID MDI's. For raw scores, the higher the item set at which testing began, the higher the obtained scores. When testing began at the item set closest to the child's CA, results were skewed in the direction of obtaining a high score
- McConachie, Couteur, & Honey, 2005 Young children suspected of having ASD were followed over time to document specific tools' abilities to detect ASD at a young age. The Mullen was chosen because of its ability to assess using a small number of items. Results indicated that only three children met the criteria for ASD before age 3. It was later determined that 79 of these children met the criteria for ASD.
- McLean, McCormick, Bruder, & Burdg, 1987 High concurrent validity, interrater reliability and internal consistency for the BDI. Favorable results for younger age group.
- Merrell & Mauk, 1993 Weak to modest correlations between the BDI and SSRS. The authors caution that the BDI has limited evidence for predicting social-behavioral development.
- Mott, 1987 The BDI did not have statistically significant correlations with any of the other language measures for the receptive communication sub domain. The author suggests the BDI may be an appropriate tool for measuring speech and language for ages 3 to 5.
- Ottem, 2003 No clinical data for WPPSI-R. On the WPPSI, the scoring level on the processing-dependent factor "knowing how" was sig. lower compared to the knowledge-dependent factor "knowing that", and the scoring level on the processing-dependent factor "seeing how" was sig. lower compared to the knowledge-dependent factor "seeing that."
- Provost, Heimerl, McClain, Kim, Lopez, & Kodituwakku, 2004 High to very high correlations found between age-equivalent scores for the BSID-II Motor scale and Peabody Developmental Motor-2 subscales. Moderate to high correlations found between the BSID-II PDIs and the PDMS-2 quotient scores. All children who scored very poor on the PDMS-2 also scored sig. delayed on the BSID-II.
- Ross, 1985 Both premature and full-term infants' scores in the normal range on the BSID MDI and PDI. Full-term infants scored sig. better on both the Mental and Motor scales than premature infants. All children scored sig. higher on the Mental scale than the Motor scale. Premature infants demonstrated greater differences between Mental and Motor scales than full-term children. Premature infants showed greater variability than full-term infants on the Mental and Motor scales.
- Saylor, Boyce, Peagler, & Callahan, 2000 At age 3 the SBIV correctly identified only 13% of the children found by the BDI to be "delayed". The SB IV and the BDI demonstrated good co-negativity at both one and two SDs below the mean.
- Sexton, McLean, Boyd, Thompson, & McCormick, 1988 BDI was able to discriminate skills of young children with disabilities.
- Snyder, Lawson, Thompson, Stricklin, & Sexton, 1993 The authors warn that BDI users should be cautious about using the BDI to obtain and report isolated performance scores.
- Tingey, Mortensen, Matheson, & Doret, 1991 BDI and Cattell are highly correlated when used to assess young children with Down syndrome.