Brief Report: Cautions Against Using the Stanford-Binet-IV to Classify High-Risk Preschoolers

Conway F. Saylor, MD, Glenna C. Boyce, MD, Sandy M. Peagler, EdS, and Serena Ashmore Callahan, EdS

The Citadel; Early Intervention Research Institute, Utah State University; and Temple University

Objective: To explore concurrent and predictive validity of the Stanford-Binet: Fourth Edition (SB-IV) by comparing scores on the SB-IV with scores from the Battelle Developmental Inventory (BDI) and later achievement scores in preschoolers at risk due to very low birthweight, and/or intraventricular hemorrhage (IVH) and other medical complications.

Methods: At ages 3, 4, and 5, 92 preschoolers were tested with the SB-IV and BDI as part of an 8-year early intervention follow-up.

Results: The SB-IV and BDI concurrent correlations at ages 3, 4, and 5 were statistically significant ($r = .73-.78, p < .0001$), as were predictive correlations ($r = .58-.85, p < .0001$). However, the BDI and SB-IV failed to place the children in the same categories for intervention services. With the BDI as the comparison measure, SB-IV failed to detect 87% of the children who were “delayed” (by BDI) at age 3 and 50% of the “delayed” children at age 5.

Conclusions: Caution is recommended when using the SB-IV to assess high risk for early intervention eligibility.

Key words: Stanford-Binet:IV; NICU; preschool assessment; intraventricular hemorrhage; early intervention eligibility.

Pediatric psychologists are increasingly called on to participate in interdisciplinary teams that follow high-risk children after their discharge from a neonatal intensive care unit (NICU). The majority of the infants discharged from tertiary care center NICUs were born prematurely and are at risk for developmental delays due to low birthweight and the medical complications that accompany it. Whether initially collected for clinical or research purposes, the psychologist’s assessment data may be utilized by families, referring physicians, and early intervention service providers to determine the child’s service eligibility. As experts in psychometrics, psychologists must take the lead in critically evaluating the instruments used with this high-risk clinical population.

The Stanford-Binet Intelligence Scale, 4th ed. (SB-IV; Thorndike, Hagen, & Satler, 1986) has been widely accepted and included among tests endorsed for assessment of children’s cognitive ability for preschool children, in spite of a shortage of studies ex-
amining its validity with younger populations. Sattler (1990) contends that the SB-IV is one of the best intelligence tests available because it has been well normed and has excellent reliability and validity in the general population. Concurrent validity studies suggest the SB-IV is likely to yield composite scores similar to those provided by other acceptable measures of cognitive functioning, like the WISC-R, WAIS-R, and Form L-M (Hollinger & Baldwin, 1990; Sattler, 1990). However, Flanagan and Alfonso (1995) urge caution when selecting the SB-IV as a measure of intelligence for preschoolers, especially with those who may be developmentally delayed, as the subtests on the SB-IV recommended for children age two or three have inadequate floors. In fact, only at the 5-year age level do all the recommended subtests have acceptable floors (Flanagan & Alfonso, 1995).

This study examined the validity of the SB-IV for classifying NICU graduates assessed during their preschool years for developmental delays. Examining the relationship between the SB-IV and more in-depth developmental evaluations in a clinical sample followed longitudinally enabled us to examine the SB-IV’s validity both concurrently (agreement and correlation with other measures given at the same time) and predictively (correlation with intelligence and achievement measures 3–5 years later).

Method

Participants

The participants were 92 preschool children who experienced intraventricular hemorrhage (IVH), and/or birthweight less than 1000 grams, along with other medical complications, secondary to premature birth. All children had been patients in the NICU and were being followed longitudinally as part of a larger study of early intervention effectiveness (for details on studies and subject characteristics see Saylor, Casto, & Huntington, 1996; Boyce, Smith, & Immel, 1993).

Attrition due to both family and project variables led to different numbers of children being sampled each year, ages 3–8. Of the 92 children testable on the SB-IV at age 3 (this excluded 17 from the original sample who were untestable due to motor or sensory impairment), 82 returned for testing at year 5. In year 4, only 72 were tested due to staffing changes and diminished resources, which led to 10 children being temporarily lost to follow-up. In year 7, 75 returned for testing on the Woodcock Johnson Test of Achievement-Revised (WJ-R). As funding ran out halfway through the eighth year of the project, only 60 children had reached their eighth birthdays in time to be tested. Analyses reported elsewhere (Boyce, Saylor, & Alexandrova, 1996) showed that the children who stayed throughout the project were comparable medically and demographically to the original sample.

Measures and Procedures

The 3-, 4-, and 5-year follow-up assessments of early intervention participants included the Battelle Development Inventory (BDI; Newborg, Stock, Wnek, Guidubaldi, & Svinick, 1984) and the SB-IV (Thordike, Hagen, & Sattler, 1986), administered within an hour of each other by master’s level clinical child/pediatric psychology assistants who were trained to 95% reliability on test administration. At ages 6, 7, and 8 age-appropriate subscales of the WJ-R (Woodcock & Johnson, 1989) measured skills and acquired knowledge. Subjects returning at age 8 also were administered the Wechsler Intelligence Scale for Children-Revised (Wechsler, 1974). At the time of testing, this study was not anticipated, so order of tests administered was not systematically recorded or counterbalanced.

Results

Correlations Among Measures

Table I presents the Pearson product-moment correlation coefficients between the SB-IV at ages 3, 4, and 5 with concurrent and future developmental and achievement scores. Concurrent analyses of the SB-IV Composite and the Battelle Developmental Quotient revealed statistically significant correlations ($r = .73$–$.78$, $p < .0001$), suggesting the two measures yield similar findings for high-risk preschoolers ages 3 to 5. In addition, the predictive validity analyses found significant correlations between SB-IV (at 3, 4, and 5) and later achievement and intelligence scores (at ages 7 and 8) ranging from .58 to .85 ($p < .0001$). Correlations were also examined between the BDI and other measures of intelligence and achievement, concurrently and
in the co-positivity from age three to age five, but even at age five, the SB-IV was only able to identify 50% of high-risk children with DQs less than 70 on the BDI (see Table II).

Discussion

This study examined the validity of the SB-IV for classifying preschoolers at risk for developmental delays. Concurrently, there were strong positive correlations between the SB-IV and the BDI. However, further investigation revealed they did not classify “high-risk” children in similar intervention eligibility categories. With the BDI as the reference standard, the SB-IV failed to identify any children with DQs less than 55 until five years of age and only agreed that 3 of the 23 “delayed” children were service eligible. This supports findings from Flanagan and Alfonso (1995) that at early ages the SB-IV subtest floors will not go low enough to document delays.

Co-positivity and Co-negativity

With the BDI serving as the reference standard, the high-risk children were classified as “delayed” or “not delayed” on both SB-IV and BDI. In the first set of analyses, “delay” was operationalized as one standard deviation (SD) below the mean, while in the second set, we used the more rigorous cut-off of two SDs. These cut-offs were chosen because, while some states use a 1.5 SD cut-off for early intervention services, others require 1 SD in two or more domains or 2 SDs in one domain. The SB-IV and BDI demonstrated good co-negativity at both one and two SDs (1 SD = 15 pt., 2 SD = 30 pt.) below the mean. That is, 100% of the children deemed “not delayed” by the BDI (score ≥ 70) were found to be not delayed on the SB-IV. When a more moderate cut-off (≥85) was used, more than 85% of the “non-delayed” children were identified as not delayed. In contrast, the SB-IV demonstrated extremely poor co-positivity with the BDI as the reference standard.

At age three, the SB-IV correctly identified only 13% of the children found by the BDI to be “delayed” (DQ < 70) or 44% of those found to be mildly delayed (DQ < 85). There was a modest improvement in the co-positivity from age three to age five, but even at age five, the SB-IV was only able to identify 50% of high-risk children with DQs less than 70 on the BDI (see Table II).

This finding is particularly troubling because data generally support the idea that developmental outcomes are better for children referred early (in the first 3 years) versus later (over age 3) (Casto & Mastropieri, 1986), and federal laws mandate timely identification of children needing services (e.g., Public Law 99-457, 1986).

### Table I. Summary of Correlations of SB-IV and DI at Ages 3, 4, and 5 With Concurrent and Future Test Scores

<table>
<thead>
<tr>
<th>Developmental Achievement Measures</th>
<th>Age 3</th>
<th>Age 4</th>
<th>Age 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>SB-IV 3</td>
<td>BDI 3</td>
</tr>
<tr>
<td>BDI 3</td>
<td>92</td>
<td>.73</td>
<td>1.00</td>
</tr>
<tr>
<td>BDI 4</td>
<td>73</td>
<td>.61</td>
<td>.70</td>
</tr>
<tr>
<td>BDI 5</td>
<td>84</td>
<td>.59</td>
<td>.76</td>
</tr>
<tr>
<td>SB-IV 3</td>
<td>92</td>
<td>1.00</td>
<td>.73</td>
</tr>
<tr>
<td>SB-IV 4</td>
<td>72</td>
<td>.78</td>
<td>.63</td>
</tr>
<tr>
<td>SB-IV 5</td>
<td>82</td>
<td>.69</td>
<td>.67</td>
</tr>
<tr>
<td>WJ Skills 6</td>
<td>77</td>
<td>.62</td>
<td>.66</td>
</tr>
<tr>
<td>Broad 6</td>
<td>77</td>
<td>.65</td>
<td>.66</td>
</tr>
<tr>
<td>WJ Skills 7</td>
<td>75</td>
<td>.67</td>
<td>.65</td>
</tr>
<tr>
<td>Broad 7</td>
<td>75</td>
<td>.68</td>
<td>.63</td>
</tr>
<tr>
<td>WJ Skills 8</td>
<td>60</td>
<td>.65</td>
<td>.66</td>
</tr>
<tr>
<td>Broad 8</td>
<td>60</td>
<td>.62</td>
<td>.64</td>
</tr>
<tr>
<td>WISC-R 8</td>
<td>55</td>
<td>.70</td>
<td>.74</td>
</tr>
</tbody>
</table>

All scores correlated significantly at \( p < .0001 \).

BDI = Battelle Developmental Inventory Score, BDI 3 = BDI, at age 3; BDI 4 = BDI, at age 4; BDI 5 = BDI, at age 5; SB-IV = Stanford Binet Intelligence Scale, Fourth Edition; SB-IV 3 = SB-IV at age 3; SB-IV 4 = SB-IV at age 4; SB-IV 5 = SB-IV at age 5; WJ = Woodcock Johnson Test of Achievement-Revised; Skills = WJ-R Skills Knowledge Domain Score; Broad = WJ-R Broad Knowledge Domain Score.
This study shows how two measures can be highly correlated (suggesting good concurrent validity) and can both correlate with future outcomes (suggesting good predictive validity) but give different clinical dispositions. In this example, using the SB-IV as the basis for early intervention eligibility (compared to the BDI) would have yielded an unacceptably high rate of underreferral of high-risk NICU graduates at preschool ages.

The SB-IV's preschool age floor problems should be assessed against other criteria besides the BDI and in other populations besides NICU graduates. The absence of a universal "gold standard" for measuring preschoolers' development and the potential for intervention placement and alternative "criterion" test scores to be influenced by one another make it difficult to arrive at the best test for high-risk preschoolers. However, the bottom line in this multisite sample of preschool children with known perinatal risk factors is that 87% of the 3-year-olds eligible for intervention services based on one test would have been ineligible if SB-IV scores had been required for placement.

This study suggests that the SB-IV, though useful for other clinical purposes, may not be a good test for evaluation of early intervention candidates, compared to other tests. In high-risk populations, the SB-IV may miss a large percentage of low functioning preschoolers who might benefit the most from early intervention programs.

Acknowledgments

This work was supported by funds from the U.S. Department of Education awarded to Utah State University's Early Intervention Research Institute (contract HS90010001). We thank Kim Foster, Sherri Stokes, Judy Moore, Sandy Glover, Kat North, Cora Price, Glen Casto, Karl White, Mark Innocenti, and Diane Behl for their participation and consultation.

Received March 13, 1998; revisions received October 19, 1998, and May 20, 1999; accepted June 10, 1999

References


Newborg, J., Stock, J., Wnek, L., Guidubaldi, J., & Svinicki,


