STABILITY OF THE BAYLEY MENTAL SCALE OF INFANT DEVELOPMENT WITH HIGH RISK INFANTS

Alison Niccols and Andrew Latchman

Introduction

The Bayley Scales of Infant Development (BSID; Bayley, 1969) have been widely used in clinical assessments and research on infants (Sattler, 1992). The Mental scale provides an assessment of sensory-perceptual abilities, object constancy, memory, learning and problem solving ability, communication and verbal skills, and early abstract reasoning ability (Whatley, 1987). It was standardised on a representative sample of 1262 American children from 2 to 30 months of age.

The Bayley Scales of Infant Development Second Edition (BSID-II; Bayley, 1993) represents the first revision of this test. The BSID-II manual reports a correlation of 0.62 between BSID and BSID-II Mental Development Index (MDI) scores from Mental Scale counterbalanced administrations to 200 children. This moderate correlation was attributed to the changes made to the BSID for the new version (i.e., expanded and improved content and coverage, extended age range, updated norms including children with clinical diagnoses; Bayley, 1993). With only a moderate correlation between the two Scales, one cannot assume the results of studies on the BSID are reflective of the BSID-II (Nellis and Gridley, 1994). Relatively few studies have been conducted to examine the psychometric properties of the BSID-II.

The revisions of the BSID were intended to improve its clinical utility. However, Nellis and Gridley express reservations about the application of Bayley scores for classification of young infants with developmental delays, due to the steep item gradients in the lower end of age levels, which may result in children being viewed as more or less delayed than they truly are (Nellis and Gridley, 1994). Nonetheless, in their study of concurrent
administrations of the BSID and BSID-II to 49 premature infants, Goldstein et al. (1995) found that, when MDI scores for BSID and BSID-II were grouped into three categories (1 standard deviation below normal, 1-2 standard deviations below normal and 2 standard deviations below normal), there was very high agreement between the two tests (kappa = 0.85) (Goldstein et al., 1995). Although the BSID-II was expected to have a major impact on the infant-toddler assessment field, no studies have examined the stability of this test with high risk infants.

**Stability of Test Scores Over Time**

The BSID and BSID-II test developers contend that prediction of future intelligence is not the purpose of their tests, however many continue to view Bayley results as predictive of later development. With typically developing infants, infant intelligence tests do not reliably predict later cognitive functioning (Fagan and Singer, 1983; Goffeney et al., 1971; Illingsworth, 1971; Lewis and McGurk, 1972; Molfese and Acheson, 1997). In general, there is an increasing stability of scores with increasing age (Molfese and Acheson, 1997; Bernheimer and Keogh, 1988; Cyphers and Fulker, 1990). In populations at developmental risk, more consistency has been found between assessments. Specifically, for children with developmental delay, the BSID is considered a good predictor of cognitive functioning later in infancy and childhood (Bernheimer and Keogh, 1988; Cook et al., 1989; DuBose, 1976). In the study reported on here, we examine stability from the first to second year of life in high risk infants.

Studies examining the stability of test scores of infants with developmental delay typically examine heterogeneous groups. Children with two of the most common diagnostic reasons for referral to infant development programmes (Down syndrome and medical conditions) generally have been considered together in these studies. Yet these two groups of children may develop differently, and the stability of the BSID-II may not be ubiquitous in delayed populations. In fact, studies of mental development in children with Down syndrome have found that BSID scores drop significantly over the first two years of life (Carr, 1970), whereas studies of medically fragile children have shown that, when corrected ages are used, Bayley scores may be fairly consistent across the first two years of life, increasing or decreasing a few points, depending on the sample examined (Rose et al., 1991; Singer et al., 1997; Sostek et al., 1987; Wilson, 1985). Therefore, if the revision of the BSID is sensitive to developmental issues in high risk samples (as it was intended to be), these patterns should be evident when the BSID-II is used with these children.

**The Present Study**

In the present study, the BSID-II was compared with the BSID in terms of its use with atypical populations to see if expected developmental patterns of scores are obtained for two samples of high risk infants. Specifically, the following issues were addressed, both of which are substantive and methodological: (a) the stability of Bayley scores for specific subpopulations of high risk infants, and (b) the stability of scores on the new Bayley in comparison to the old Bayley. The hypotheses to be tested were:
Hypothesis 1: For infants with Down syndrome, scores will not be stable from the first year of life to the second year of life.

Hypothesis 2: For medically fragile infants, scores will be stable from the first year of life to the second year of life.

Hypothesis 3: Both the BSID and the BSID-II will show the expected pattern of stability from the first to the second year of life for children with Down syndrome.

Hypothesis 4: Both the BSID and the BSID-II will show the expected pattern of stability from the first to the second year of life for medically fragile children.

We predicted that, for children with Down syndrome, scores obtained on either the BSID or the BSID-II in the first year of life would be higher than scores obtained in the second year of life. For medically fragile children, scores obtained on either the BSID or the BSID-II in the first year of life would not be different from scores obtained in the second year of life. These predictions derive from the general hypothesis that any instability in scores would be due to the nature of infant development in specific groups, rather than deficiencies in the tests.

Method

Subjects

Subjects were 66 infants referred to infant development/early intervention programmes due to a diagnosis of Down syndrome or multiple medical conditions (medically fragile). Thirty-two (48%) of the infants were diagnosed with Down syndrome, thirty-four (52%) were medically fragile. All infants resided in urban/suburban areas of Southern Ontario, Canada: Thirty-three of the infants were from the Hamilton-Wentworth, 28 from London, 2 from Brampton and 3 from Waterloo. Thirty-five infants (53%) were male, and fifty-six (85%) came from two-parent families. Each infant in the medically fragile group had multiple medical conditions: Forty-four percent were born prematurely (birth at less than 37 weeks gestation; Goldstein et al., 1995), and other medical conditions included central nervous system difficulties (e.g., seizure disorder), heart conditions/defects (e.g., patent ductus arteriosus), respiratory problems (e.g., bronchopulmonary dysplasia), feeding problems (e.g., swallowing disorder), organ malfunctions (e.g., hydronephrosis), skeletal disorders (e.g., osteoporosis), metabolic disorders (e.g., hyperglycemia), and infections (e.g., pneumonia). Infants were excluded from the study if they had an identified genetic disorder (other than Down syndrome), if there was known prenatal exposure to drugs or alcohol, or if their only presenting problem was parenting concerns, cerebral palsy, or developmental delay (with no Down syndrome or medical conditions).

Procedure

After receiving approval from ethical review boards in each of the agencies, psychological files and health records were examined and assessed for applicability to the study. Data for the BSID-II sample and the BSID sample were derived from files of infants who completed one of the tests between 4 and 10 months of age (mean age at the first assessment = 7, SD = 2), and the same test again between 17 and 25 months of age (mean age at the second assessment = 22, SD = 2), and who had
both tests administered by a trained psychometrist. The 33 infants who completed the BSID-II from 1994 to 1997 were matched with 33 infants who completed the BSID from 1987 to 1994 by primary diagnosis (Down syndrome or medically fragile), age at the first assessment, age at second assessment, gender, and geographic region in which the test was administered. All 33 matches had the same primary diagnosis, 28 were from the same region, 29 were matched on gender and 31 were matched within two months of age at assessment. For one of the matches, the age differential was 3 months at the first assessment, and for one other match the age differential was 6 months at the second assessment. None of the matches exceeded two months of age difference at both the first and second assessments. There were no significant age differences between the group who completed the BSID-II and those who completed the BSID. A 3-way analysis of variance (ANOVA) with two Between Groups factors (Down syndrome vs medically fragile and BSID vs BSID-II) and one Within Groups factor (Time 1 vs Time 2) revealed an interaction of Time by Diagnostic group, F(1, 62) = 17.45, p < .001, on chronological ages. The children with Down syndrome were significantly younger than the medically fragile children at the first assessment, t(60.23) = -2.51, p < .05 (mean age = 7 vs. 8 months), and older than the medically fragile children at the second assessment, t(64) = 3.48, p < .001 (mean age = 23 vs. 21 months). [Similar findings were obtained when analysing corrected ages. In terms of corrected ages, the children with Down syndrome were significantly older than the medically fragile children at the second assessment only, t(58.87) = 4.50, p < .001 (mean age = 23 vs. 20 months)].

Additional information recorded from the files was: degree of prematurity, medical complications (Hamilton only), number of parents, raw scores for the Mental Scale, and Mental Development Index (MDI) scores. For statistical analyses, Mental Development Index scores were not used because many children had unspecified scores “under 50”. Instead, Developmental Quotients (DQ) were calculated using the traditional mental age method (Mental Age/Chronological Age x 100) with chronological ages always corrected for prematurity. The mean Developmental Quotient (DQ) for the 32 infants with Down syndrome who completed the BSID or BSID-II was 82 (SD = 20) in the first year of life (which corresponds to an IQ level in the “low average” range; Grossman, 1983) and 66 (SD = 14) in the second year of life (which corresponds to an IQ level in the “intellectually deficient” range; Grossman, 1983). The corrected mean DQ for the 34 infants with multiple medical conditions was 71 (SD = 23) in the first year of life and 70 (SD = 23) in the second year of life.

**Results**

A 3-way analysis of variance (ANOVA) with two Between Groups factors (Down syndrome vs medically fragile and BSID vs BSID-II) and one Within Groups factor (Time 1 vs Time 2) was used to analyse the DQ scores. Results of this analysis revealed a significant three-way interaction of Diagnostic group by Bayley version by Time, F(1,62) = 4.77, p < .05. [Analyses controlling for T2 age produced similar findings, ANCOVA F(1,60) = 4.20, p < .05.] Mean DQ scores from the BSID decreased more than one Standard Error of the Mean (SEM) from the first to second year for
Mean DQ scores from the BSID-II decreased more than one SEM from the first to second year for the children with Down syndrome, and increased more than one SEM from the first to second year for the medically fragile children. The performance of children with Down syndrome from the first to second year was consistent across Bayley versions (both revealing lower scores in the second year), but the performance of medically fragile children from the first to second year differed across Bayley versions (with BSID scores lower in the second year than the first, and BSID-II scores higher in the second year than the first). (See FIGURE 1.) In order to examine stability over time, correlations between first year DQs and second year DQs were computed by diagnostic category and Bayley version. Results revealed significant correlations for children with Down syndrome, $r = .51, p < .05$ for BSID and $r = .65, p < .01$ for BSID-II. For medically fragile children, $r =$.
.67, p < .01 for BSID and r = .37, n.s. for BSID-II. These correlations indicate moderate stability (or moderate systematic change) from the first to second year for these tests when used with high risk infants.

To examine classification agreement over time, BSID and BSID-II MDIs were divided into groups that represented Borderline or Normal performance (-2 SD and above) and Abnormal performance (below -2 SD) for each of the two diagnostic groups at each assessment. For infants with Down syndrome, many obtained MDI scores within the Abnormal range in their second year but not their first year, and classification agreement (Kappa) was low, as expected, for both the BSID and the BSID-II. For infants with multiple medical conditions, agreement between first and second year MDI classifications was moderate for both the BSID and the BSID-II. (See TABLE I.)

**Discussion**

For the children with Down syndrome in the present study, mean DQ scores on the BSID-II obtained in the first year of life were comparable to scores found in other studies using the Bayley (Saxon and Witriol, 1976; Wishart and Duffy, 1990). In the second year of life, study sample scores were similar to scores reported for the 60 children with Down syndrome in the BSID-II standardisation sample. Mean DQ scores for the children with multiple medical conditions in the sample were similar to scores reported by Bradley and Casey (1992) and Rothberg et al. (1981) and to scores reported for the 35 children who were HIV positive (and experiencing a range of medical complications) in the BSID-II standardisation sample (Bayley, 1993). Mean DQ scores for the medically fragile children in the study sample were lower than scores reported for some other samples of children with medical conditions (who were presumably less impaired) (Rose et al., 1991; Singer et al., 1997). These replicated results lend support to the clinical validity of the BSID-II. As stated in the BSID-II manual, “the BSID-II is sensitive to performance differences between... children with various conditions placing them at risk for delayed development.” (Bayley, 1993, p. 225)

The results of this study indicated a significant interaction of group (diagnostic category), time (age at assessment), and Bayley version. Put simply, mean DQ

**TABLE I**

<table>
<thead>
<tr>
<th>Group</th>
<th>BSID</th>
<th>BSID-II</th>
<th>BSID</th>
<th>BSID-II</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Down syndrome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSID</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>.03</td>
</tr>
<tr>
<td>BSID-II</td>
<td>4</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>.04</td>
</tr>
<tr>
<td>Medically fragile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSID</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>.54</td>
</tr>
<tr>
<td>BSID-II</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>.41</td>
</tr>
</tbody>
</table>

Note: Children with MDI scores <70 in both the first and second year are coded as “++.”
Those with MDI scores <70 in the first year and ≥70 in the second year are coded as “+-.”
Children with scores ≥70 in the first year and <70 in the second year are coded as “-+.”
Those with scores ≥70 in both the first and second year are coded as “--.”
scores dropped 13-18 points from the first to second year for children with Down syndrome assessed with the BSID or the BSID-II and for medically fragile children assessed with the BSID. For medically fragile children assessed with the BSID-II, however, the mean DQ score increased 11 points from the first to second year of life. These findings reflect an interaction of sample/population, developmental, and test issues.

Although infant mental development may appear stable in heterogeneous groups of high risk infants (Bernheimer and Keogh, 1988; Cook et al., 1989; DuBose, 1976), there may be expectable changes for specific subgroups such as infants with Down syndrome and those who are medically fragile. For infants with Down syndrome, mental development appears to decelerate over infancy and early childhood, regardless of the test used (Cicchetti and Beeghly, 1990). Some have speculated that this happens when there is a shift from subcortical to cortical functions (Saxon and Witriol, 1976), others have suggested that test items for young infants tap perceptual skills that may be relatively intact in Down syndrome whereas items for older infants tap communication skills that may be less well developed in Down syndrome (Cicchetti and Beeghly, 1990; Kopp, 1983); and others have suggested that the gap between the performance of typically developing children and those with Down syndrome simply widens over time (Cicchetti and Beeghly, 1990). Whatever the reason, the infants with Down syndrome who were assessed with the BSID-II in the present study had DQ scores that dropped from the first to second year, as has been found in previous studies, and, indeed, in the study sample of children with Down syndrome who had completed the BSID. Correlations between first and second year scores were significant and indicated a moderate level of systematic change. Classification agreement over time was low, as expected, with most children with Down syndrome falling in the Abnormal category at the second testing but not the first. These findings lend further evidence of the validity of the BSID-II, as they reveal the expected pattern of early mental development in children with Down syndrome.

For the medically fragile infants who were assessed with the BSID-II in the study, DQ scores increased from the first to second year of life, perhaps as their medical conditions improved. Results from medically fragile infants who completed the BSID were somewhat different, showing a decrease in scores from the first to second year. Although the pattern of results obtained for the BSID and the BSID-II differed, they were both somewhat similar to previous research on infants with multiple medical conditions showing that mental development usually is impaired to some extent, with a slight increase or decrease in test scores over the first two years of life depending on the sample examined (Rose et al., 1991; Singer et al., 1997; Sostek et al., 1987; Wilson, 1985). The differential pattern of results over time for the BSID and the BSID-II when used with medically fragile infants could be interpreted as evidence for questioning the validity of either (or both) test(s), but it is more likely a result of group differences. Despite matching samples on primary diagnosis, age at the first assessment, age at second assessment, gender, and geographic region in which the test was administered, groups may have differed on other factors (e.g., psychosocial factors, number, type, severity, combination, and/or management of medical illnesses, all of which may be associated with differential cognitive outcomes) (Sostek et al., 1987;
Nevertheless, although BSID-II scores in the first year of life may underestimate cognitive abilities in the second year of life for medically fragile children, correlations between first and second year scores indicated a moderate level of stability (or systematic change), and classification agreement over time was moderate. Thus, for medically fragile infants, developmental delays identified in the first year with the BSID-II may indicate a moderate probability of continuing delays. The BSID-II appears to be sensitive to developmental changes in the first two years of life of infants with Down syndrome and those with multiple medical conditions. The findings of this study support the clinical validity of the BSID-II. While the BSID-II as a measure of infant development may be a valid and useful indicator of current functioning, scores for some high risk infants, especially those with Down syndrome, may not be stable from the first to second year of life. To improve prediction of second year scores from first year scores for infants with Down syndrome, future research with a larger sample could focus on calculating a regression slope, and identifying factors accounting for the variance in prediction.

Although some consistency has been found with respect to mental development in children with developmental delay across infancy (Cook et al., 1989; Illingsworth, 1961; VanderVeer and Schweid, 1974), the present study indicates that this stability is not ubiquitous in delayed populations. In this study, BSID-II DQ scores for two of the largest groups of high risk infants referred to infant development programmes (those with Down syndrome and those with multiple medical conditions) differed, as did the pattern of their performance from the first to second year of their lives. These findings have clinical implications and are directly relevant to assessment policies and practices in infant development programmes. Together with other findings regarding mental development in specific groups of high risk infants, they invite cautious interpretation of assessments conducted in the first two years of life. Specifically, results from BSID-II assessments of high risk infants in the first year of life should not be used for predictive purposes, and must be interpreted differently for infants with Down syndrome than for infants with multiple medical conditions.

There were several limitations to interpretation of the results of this study that are important to consider. First, although the best available indicator of mental development in children with developmental delays, DQ, was used, the validity of the DQ as an indicator of cognitive development has been questioned (Sattler, 1992; Colombo, 1993). Developers of future instruments should include more potentially handicapped children in their standardisation sample so that MDIs below 50, and even below 30 (Robinson and Mervis, 1996), may be calculated. Second, for the BSID DQ calculations, the ratio of raw to standard scores in the mental age conversion chart is lower than that for the BSID-II, rendering the DQ calculations for the BSID less accurate, and comparisons between DQs for the two tests problematic. Third, different examiners administered the tests (i.e., there were different psychometrists in different regions), perhaps introducing variability in administration (e.g., item set selection), which can influence scores (Black and Matula, 2000; Gauthier et al., 1999; Ross and Lawson, 1997; Washington et al., 1998). Fourth, variability was likely increased also by heterogeneity in the
medically fragile group, relatively wide age ranges, and differences in age cohorts. Fifth, sample size precluded within group comparisons.

The findings of this study suggest that specific groups of high risk infants may perform differently over the first two years of life on the BSID-II. Future research examining the mental development of other specific groups of high risk infants would be informative to the field of infant development and to clinical early intervention programmes.

Summary

This study examines stability of scores on the Bayley Mental Scale of Infant Development Second Edition (BSID-II) across the first two years of life for high risk infants. BSID-II scores in the first and second year of life for 16 infants with Down syndrome and 17 medically fragile infants were compared to Bayley Mental Scale of Infant Development (BSID) scores for 33 matched infants. A three-way interaction of Group, Time, and Bayley Version on developmental quotient (DQ) scores revealed that, for infants with Down syndrome, scores decreased from the first to second year for both test versions, but for medically fragile infants, BSID scores decreased from the first to second year and BSID-II scores increased from the first to second year. These results indicate that the BSID-II is sensitive to patterns of developmental changes in the first two years of life that are specific to infants with Down syndrome and to medically fragile infants. Together with other findings regarding mental development in specific groups of high risk infants, they invite cautious interpretation of assessments conducted in the first two years of life. Specifically, results from BSID-II assessments of high risk infants in the first year of life should not be used for predictive purposes, and must be interpreted differently for infants with Down syndrome than for infants with multiple medical conditions.

Acknowledgements

This study was financially supported by the Infant-Parent Program, Chedoke Child and Family Centre, Children’s Hospital, Hamilton Health Sciences and McMaster University. This paper was submitted in partial fulfillment of Andrew Latchman’s B.Sc. degree in Bio-Psychology at McMaster University. This research was presented at the 1999 Society for Research in Child Development Conference, Albuquerque, NM, and the 1999 Ontario Association for Infant Development Conference in Mississauga, Canada.

We are grateful to the Infant Development Programs (Hamilton-Wentworth, London, Peel, and Waterloo) who allowed us to access their records to conduct the study. We thank Leslie Atkinson for his help in planning the study, Rosalina De Sa for her help with data collection, Helen Williamson for her help with data analyses, and Michael Boyle, Charles Cunningham, Kathleen Kitching, Michelle Van Tuyl, Helen Williamson, and anonymous reviewers for their helpful comments on earlier drafts of the manuscript.

References


Relation to birth status and developmental outcome during the first 5 years. Developmental Psychology, 27, 723-737.


