0196-206X/99/2002-0075
DEVELOPMENTAL AND BEHAVIORAL PEDIATRICS

Vol. 20, No. 2, April 1999
Printed in U.S.A.

Original Articles

The Bayley Scales of Infant Development II: Where to Start?

SUSAN M. GAUTHIER, M.A.
CHARLES R. BAUER, M.D.
DANIEL S. MESSINGER, PH.D.
JANINE M. CLOSIUS, B.S.
Division of Neonatology, Department of Pediatrics, University of Miami School of Medicine, Miami, Florida

ABSTRACT. This study examines whether the Bayley Scales of Infant Development-Second Edition (Bayley II) Mental Scale scores vary on the basis of which item set is considered the starting point of an infant's assessment. The Bayley II was administered to 78 12-month-old infants by certified examiners beginning with the 12-month age item set. A second certified examiner then administered 10 additional items that completed the 11-month and 13-month age item sets. Of the 78 infants tested, 73 (94%) met basal and ceiling criteria in all three item sets. Three separate Mental Developmental Index (MDI) scores were calculated using 12-month norms for each subject, which were based on the raw scores generated from the 11-, 12-, and 13-month age item sets. Scores calculated on the 11-month set were significantly lower than those on the 12-month set, which were in turn significantly lower than those on the 13-month set. When tested on the 11-month instead of on the 12-month item set, twice as many infants received lower than normal (<85) MDI scores, indicating an impact on referability decisions. Minor adjustments in administration of the Bayley II, such as those based on assumptions regarding an infant's current level of functioning, significantly affect infant test scores and eligibility for follow-up services. Standardized use of this test should minimize variability in test scores. *J Dev Behav Pediatr 20:75–79, 1999.* Index terms: *administration, Bayley II, bias, development, mental.*

Part of the process of standardizing infant testing includes addressing the possibility that examiner bias may influence test scores. This study raises the possibility that examiner bias may inadvertently alter Mental Developmental Index (MDI) scores of 1 year olds on the Bayley Scales of Infant Development-Second Edition (Bayley II).[1] This may occur if an examiner adjusts the starting point of a Bayley II assessment on the basis of his or her impression of an infant's likely performance.

In a clinical setting, standardized measures of infant ability are primarily used to assess infants who show or are suspected of showing developmental delay. As a result, these evaluation tools of infant development need sufficient reliability and validity for professionals to make confident decisions regarding classification, service levels, and effectiveness of intervention methods. Until recently, the Bayley Scales of Infant Development (1969)[2] (BSID) were the most widely used measures of infant development. They were adopted by clinicians as a guide for the diagnosis of developmental delay and placement into early intervention services, as well as by researchers conducting outcome studies.

Address for reprints: Susan M. Gauthier, M.A., Department of Pediatrics, University of Miami, P.O. Box 016960 (M-827), Miami, FL 33101.
Portions of these data were presented at the meetings of the 1997 Southern Society for Pediatric Research, New Orleans. This research was supported by the National Institute of Child Health and Human Development (NICHD) through cooperative agreement U10 HD21397.

In the original BSID,[2] suggested starting points for infants of different ages were provided. Basal and ceiling rules involved passing and failing, respectively, 10 consecutive items. After 25 years of widespread use, concerns arose regarding potential weaknesses of the BSID such as the possibility that BSID norms were no longer reflective of the current population in the United States.[3]

In response to these concerns, the Bayley II was developed and released in 1993 using restandardized norms. Like its predecessor, the Bayley II is divided into two subscales, the Mental Scale and the Motor Scale. This study focuses on the Mental Scale of the Bayley II which is composed of 178 items of increasing difficulty. The items measure performance in the areas of sensory-perception, knowledge, memory, problem solving, and early language. In addition to becoming rapidly accepted as a thorough measure of infant functioning, the Bayley II is often used as the primary diagnostic tool when determining eligibility for early intervention services as mandated by Public Law 99-457. It is also frequently used as a developmental reference point in research settings. After fewer than 5 years on the market, new concerns have arisen regarding its own particular weaknesses, one of which is the possibility of generating "radically different scores"[4] (p. 205) if the starting point of the assessment is altered because of the assessor's suspicion of infant delay.

The basal and ceiling rules of the Bayley II differ from those of the original BSID[2] in that an infant attains a basal provided the infant successfully passes at least five items

within the item set on which he or she is being tested. The infant attains a ceiling when he or she fails a minimum of three items within that same item set. If the infant cannot successfully obtain a basal, that is, if he or she cannot pass five items, the examiner must administer the preceding item set(s) until a basal is reached. Similarly, if the infant does not achieve a ceiling, i.e., does not fail at least three items, the examiner must administer subsequent item set(s) until the infant fails three items in a set. The infant's MDI score is based on the raw score obtained on the set in which both basal and ceiling criteria are met. Additional credit is given for any items passed beyond that item set if testing was started on a higher set.

New Bayley II examiners are encouraged to participate in a training session given by The Psychological Corporation personnel. At that time, examiners are given a training manual[5] that states "It is strongly recommended that you begin the testing with the item set that is appropriate for the child's chronological age. [However,] if you have any definitive information about the child you are testing that indicates that you should test with an earlier or later item set, you may begin with an item set that does not match the child's chronological age" (p. 24). The more widely available Bayley II test manual[1] purchased with the testing kit is more lenient about where to start an assessment: "There will be times when it is desirable to begin testing with an item set that is above or below the infant's chronological age. In such cases, select the item set that you feel is closest to the child's current level of functioning based on other information you might have" (p. 42). A footnote states that "when testing a premature child under the age of 2 years, you may want to begin testing with the item set appropriate for the child's corrected age" (p. 41).

This study addresses whether changing the starting point of the Bayley II will alter an infant's test score and what implications this would have on intervention placement. The hypothesis is that 12-month-old infants tested on the Mental Scale of the Bayley II will achieve significantly different scores depending on which item set, the 11-, 12-, or 13-month, is administered.

METHODS

Subjects

Subjects consisted of 78 infants with a chronological age ranging from 11 months, 16 days to 12 months, 15 days (corrected for prematurity when necessary) who were enrolled in a longitudinal, multisite, clinical study of cocaine exposure. All of the subjects were born at a large, urban county hospital in South Florida. Of the 78 infants, 64 (82.1%) were black, 44 (56.4%) were male, and 53 (67.9%) were full-term (gestational age $\geq$ 38 weeks, birth weight $>$ 2500 g). The 53 full-term infants ranged from 38 to 41 weeks gestational age with a mean of 39.6 weeks. The 25 premature infants ranged from 24 to 37 weeks gestational age with a mean of 32.7 weeks.

Procedure

This study involved the 11-, 12-, and 13-month sets of the Bayley Scales of Infant Development-Second Edition (Bayley II) Mental Scale which encompass 40 items numbered 66 through 105 (see Fig. 1). The 11-month set includes items 66 through 92 (27 items), the 12-month set includes items 71 through 100 (30 items), and the 13-month set includes item 78 through 105 (28 items). Overall, 22 (55%) of these 40 items appear in all three item sets.

In addition to the number of items passed on a given item set, infants are given automatic credit for all items preceding the set on which they achieve a basal. For example, they automatically receive 65 base points if tested on the 11-month set, in which testing begins at item 66. Similarly, they receive 70 points if tested on the 12-month set and 77 points if tested on the 13-month set. Total base points plus the number of items passed within the item set being tested constitutes an infant's raw score. The raw scores are then translated to Mental Developmental Index (MDI) scores on the basis of the standardized norms printed in the Bayley II test manual,[1] which take into account the infant's chronological age at the time of testing (corrected for prematurity when necessary).

The Bayley II was administered to all infants by certified examiners beginning with the 12-month item set. A second certified examiner then administered ten supplementary items immediately after the completion of the 12-month item set to complete the 11- and 13-month item sets. To avoid bias in administering supplementary items, 40 subjects were first given the five lowest 11-month items (items 65 through 70) and 38 subjects were first given the five highest 13-month items (items 101 through 105).

Repeated-measures analyses of variance were performed separately on the Bayley II MDI and raw scores. Scores from the three consecutive item sets constituted the repeated measures. We also tested for the effects of order of administration, sex, race, presence of prematurity, and birth weight category. We controlled for these variables as covariates only if they had significant effects on the Bayley II score outcome variables. Follow-up comparisons between scores from adjacent item sets were conducted with repeated-measures analysis of variance between the two item sets. For analyses relevant to referability, each infant's MDI on each item set was classified as at-risk/delayed ($<$85) or normal/above normal ($\geq$ 85). The number of infants classified in both categories on the basis of their scores in each item set were examined with the Cochran Q test, which is distributed as χ^2.

RESULTS

All 78 infants tested were able to achieve basal and ceiling criteria on the 12-month item set. In addition, 73 (94%) infants achieved a basal and a ceiling in all three item sets. This is important because according to the specifications in the Bayley Scales of Infant Development-Second Edition (Bayley II) test manual,[1] all three Mental Developmental Indices (MDIs) could be acceptable scores. For the three infants who did not have an 11-month MDI because they did not obtain a ceiling on the 11-month item set, their 12-month MDI scores were substituted. There were two missing 13-month scores that occurred because the infants did not achieve a basal on the 13-month item set. These were replaced by an MDI that was based on their raw score in the 12-month item set plus those items successfully passed in the 13-month set. These substitutions follow the standard administration protocol contained in the Bayley II test manual[1] and articulated

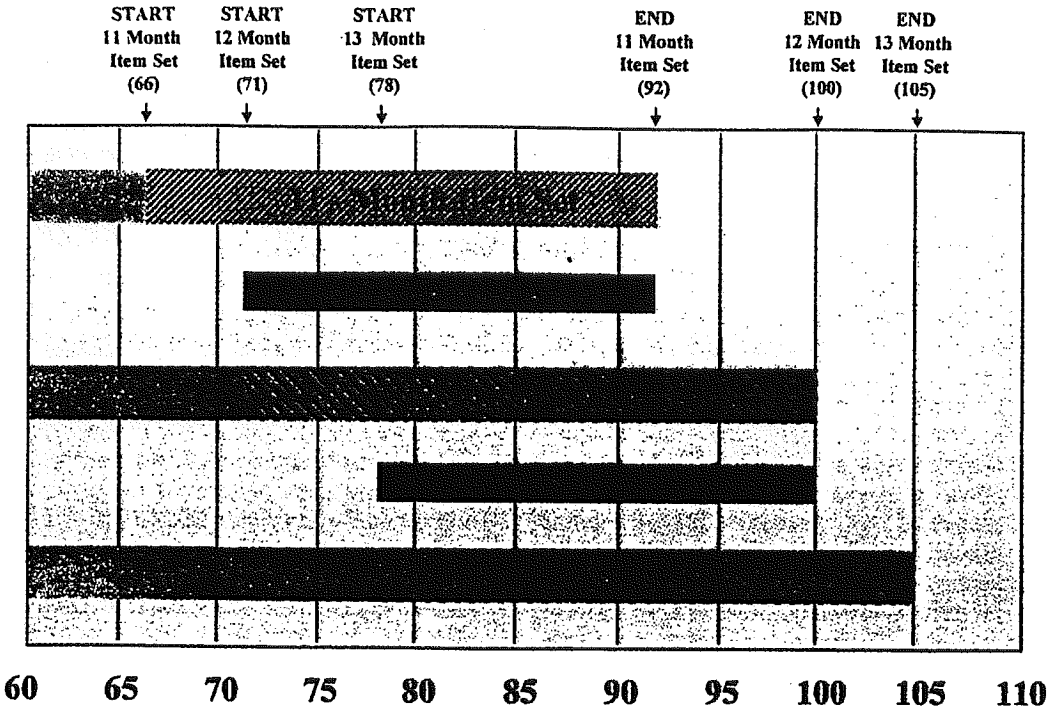

FIGURE 1. Bayley Scales of Infant Development-2nd ed. Distribution of item numbers within age item sets.

in the introduction. These procedures could potentially bias against finding a hypothesized difference because they reduced the differences between the 11- and 12- and between the 12- and 13-month scores by substituting scores equivalent to or based on the 12-month scores in these five instances.

A repeated-measures analysis of variance was performed on the infants' scores on each of the three consecutive item sets (11-, 12-, and 13-mo item sets). There were no significant main or interaction effects involving order of administration, sex, race, presence of prematurity, or birth weight category. Significant differences were found between each of the age item set raw scores (F[2,76] [Wilks'] = 241.80, $p < .0001$) as well as between the corresponding MDIs (F[2,76] [Wilks'] = 203.15, $p < .0001$) (see Table 1). Planned follow-up, repeated-measures comparisons revealed that for both raw scores and MDIs, infants scored higher on the 13-month set than on the 12-month set ($F_{raw}[1,77] = 187.47$, $p < .0001$; $F_{MDI}[1,77] = 161.49$, $p < .0001$) and higher on the 12-month set than on the 11-month set ($F_{raw}[1,77] = 323.40$, $p < .0001$; $F_{MDI}[1,77] = 226.05$, $p < .0001$).

This lack of prematurity effect indicates that the results were consistent for premature and full-term infants. Nevertheless, the analyses were run separately for the premature infants. Like the group as a whole, premature infants achieved lower MDI scores on lower age item sets. Specifically, follow-up comparisons indicated that they scored 5.6 MDI points lower on the 11-month item set ($M = 89.24$) than on the 12-month item set ($M = 94.84$) (F[1,24] [Wilks'] = 45.45, $p < .0001$); they also scored 6.16 MDI points lower on the 12-month item set ($M = 94.84$) than on the 13-month item set ($M = 101.00$) (F[1,24] [Wilks'] = 47.30, $p < .0001$).

For the sample as a whole, several factors may account for differences in raw scores and, hence, MDIs. When a higher age item set was used, infants were given automatic credit

Table 1. Bayley II Age Item Set Differences and Mental Developmental Index Data

	Age Set		
Score	11-mo	12-mo	13-mo
Raw Score			
Mean	82.78[a]	85.47[b]	87.83[c]
SD	3.48	3.85	3.87
Range	74–91	76–96	80–99
		F(2,76) (Wilks') = 241.80, $p < .0001$	
Mental Developmental Index			
Mean	88.50[a]	95.03[b]	100.78[c]
SD	9.32	9.71	9.81
Range	68–102	72–120	82–127
		F(2,76) (Wilks') = 203.15, $p < .0001$	

Bayley II, Bayley Scales of Infant Development-Second Edition; SD, standard deviation. For each variable, means with different superscripts are significantly different at $p < .0001$.

Table 2. Classification of Mental Developmental Index Score

	Age Set MDI					
	11-mo		12-mo		13-mo	
Classification	n	%	n	%	n	%
At-risk/delayed (MDI < 85)	29	37	14	18	2	3
Normal (MDI = 85–115)	47	60	62	80	73	94
Above normal (MDI > 115)	2	3	2	3	3	4

MDI, Mental Developmental Index.
The Cochran Q test (testing differences in classification between at-risk/delayed and the combined classification of normal and above normal) is distributed as χ^2. Q (2, N = 78) = 40.67, $p < .0001$. Percentages have been rounded off so totals are not equal to 100%.

for earlier items. When a lower item set was used, infants may or may not have received credit on some of these items. For example, items 66 through 70 are contained in the 11-month item set but not in the 12-month set. When infants were tested on the 12-month item set, they automatically received 5 (raw) points for these items. However, when infants were actually tested on the 11-month item set, they were only able to pass an average of 3.58 of these items. The difference between the 5 points of presumptive credit given when the 12-month item set was administered and the 3.58 points earned when the 11-month item set was administered is significant ($t[77] = 11.82, p < .0001$). These 1.42 raw points contribute to the difference between 11-month and 12-month item set MDIs. Similarly, infants whose testing began on a lower item set were not given the opportunity to attempt the highest numbered items in the next higher item set. For example, items 93 through 100 are included in the 12-month item set, but not in the 11-month set. Infants tested on the 11-month set and who met basal and ceiling rules did not attempt these 7 items and therefore received no credit, but when tested on the 12-month set, they averaged 1.27 of these items. The difference between the 1.27 earned points and the 0 points associated with testing on the 11-month set is also significant ($t[77] = 9.58, p < .0001$).

To summarize this example, the 11-month raw score was lowered by a total of 2.69 points in relation to the 12-month score. That is, if administered the 12-month set rather than the 11-month set, infants gained 1.42 presumptive points (5−3.58) plus 1.27 earned points from the higher items attempted (items 93 through 100) resulting in the 2.69-point raw score difference. These data illustrate that infants failed some of the earlier items when they were actually administered and that they passed some of the later items when given the opportunity to attempt them. This pattern of significant differences at both the low and high ends of the age item sets was also observed between the 12- and 13-month item sets. As expected, however, infants were able to pass a higher percentage of items in the 11-month set (66%) than in the 12-month set (52%) or in the 13-month set (39%).

On the basis of the standardization sample used by The Psychological Corporation, Bayley II scores are normally distributed with a mean of 100 and a standard deviation of 15. Any score within one standard deviation of the mean (between 85 and 115) is considered to fall within the normal range. Scores lower than 85 indicate at-risk or delayed functioning, and scores higher than 115 are above normal. Table 2 shows how infants' MDIs would be categorized depending on which item set was administered. The proportion of scores that fall above the normal range remain essentially unchanged regardless of which item set is administered. However, twice as many infants were categorized as at-risk or delayed when scored on the 11-month set compared with scoring on the 12-month set. Conversely, at-risk/delayed scores were virtually eliminated when the 13-month set was administered. This pattern did not differ in the prematurely born infants when contrasted with full-term infants. Among the prematurely born infants as well, 50% more (9 versus 6) fell into the at-risk category when tested on the 11-month rather than on the 12-month item set ($Q = [2, n = 25] = 12.250, p < .002$).

DISCUSSION

This study indicates that infants' scores on the Bayley Scales of Infant Development-Second Edition (Bayley II) are influenced by the item set in which testing begins. The Bayley II Test Manual[1] suggests that infants should be tested on the item set which is concurrent with their chronological or corrected age (p.41). However, it also allows for the use of clinical impression based on prior information regarding developmental progress to determine the item set with which to begin the assessment. The option of varying the starting point to a more developmentally appropriate item set, when used with infants who have an obvious severe developmental delay, would allow the tester to start at an earlier item set, thereby avoiding a prolonged testing session that could unnecessarily tire and frustrate the infant. However, our data demonstrate that simply testing infants with the Bayley II on an aged item set other than that which corresponds directly with their chronological or corrected age significantly affects their scores.

Testing prematurely born infants on the Bayley II, however, presents another dilemma. Ross and Lawson[6] found that only 4 of 28 psychologists in a limited poll would use chronological age to determine the starting point of a premature infant's assessment. Because a premature infant's corrected age is, by definition, lower than his or her chronological age, the decision of whether or not to correct for prematurity can significantly influence test scores. In fact, Ross and Lawson[6] showed that premature infants who were tested on an item set corresponding to their corrected age scored more poorly than when tested on an item set corresponding to their chronological age. Testing a premature infant at the lower corrected age does not provide the opportunity to earn points in a higher item set. In contrast, beginning the assessment with the chronological age item set allows the infant to be credited

for all items below that item set, which the infant otherwise might have failed.

In a response to a commentary by Ross and Lawson,[6] Matula et al[7] of The Psychological Corporation have responded that differences found between Mental Developmental Indices (MDIs) on different item sets, although statistically significant, are fairly minimal and have little clinical significance. The results of our study, in which 12-month norms were used consistently, cast doubts on this assertion. Among premature infants in this study, Bayley II MDI scores were lowered by a mean of 5.6 points and by as many as 15 points. Although corrected age scores were not compared with chronological age scores, it is evident that by adjusting the starting point by as little as one age item set, without adjusting age norms, clinically important test score differences were found. In fact, 50% more premature infants were classified in the at-risk to delayed category when testing was carried out using the 11-month item set instead of the 12-month item set.

More generally, all of the infants tested in this protocol were able to achieve a basal and a ceiling on the 12-month item set even though many of them were considered to be "at risk" for developmental delay on the basis of their social, economic, and medical background. However, if it were assumed, on the basis of their at-risk status, that the 11-month item set was the more appropriate starting point, the Bayley II MDI scores would have been lowered by a mean of 6.53 points, with a range up to 17 points. This exceeds one full standard deviation. In addition, twice as many infants would have been classified as either at-risk or delayed.

It has yet to be determined whether test scores will be significantly altered when testing infants beyond the 12-month age range. Nevertheless, these findings would certainly have a clinically important impact on the referral process as well as on the actual placement of infants in early intervention programs, as mandated under Public Law 99-457.

The "strong" recommendation that only "definitive" information about the infant be used in determining the starting point of the Bayley II appears only in the training manual.[5] It should be contained in the test manual[1] as well. This present inconsistency allows for unnecessary variability in administration practices. Because the Bayley II is widely used in both clinical and research settings, it is imperative that it be administered in a standardized way. This would ensure that test scores are a valid measure of an infant's level of functioning, and decisions regarding classification and placement could be made with confidence.

One way to achieve the goal of consistent administration would be to begin testing all full-term infants with the item set that corresponds to their chronological age. Examiners should only alter the item set being tested after it is demonstrated that the infant cannot achieve a basal or a ceiling. This simple directive would ensure that all test scores would be meaningful and comparable. This study was not intended to examine how correction for prematurity affects test scores. If "corrected" gestational age correlates well with chronological age, then the same recommendation might hold true for testing of premature infants. The recommendation would then be to start with the item set that corresponds to their "corrected" gestational age. Further research into the unique issues related to testing premature infants is needed to unravel this conundrum.

If infants are able to achieve the criteria for basal and ceiling on several different item sets, they may receive more than one possible test score with very different clinical implications or interpretations. If examiners are allowed to use their clinical judgment to select the starting point of the Bayley II when assessing infants for their eligibility into early intervention programs and choose to begin testing at a lower item set than that which corresponds to the infant's chronological age, more infants may qualify for services. This could have significant public policy implications regarding funding to provide services for those infants who "qualify" on the basis of these lower test scores. The far-reaching implications of these findings are yet unclear, but they demonstrate a major weakness in the psychometric properties of the Bayley II.

REFERENCES

1. Bayley N: Manual for the Bayley Scales of Infant Development, 2nd ed. San Antonio, TX, The Psychological Corporation, 1993
2. Bayley N: Manual for the Bayley Scales of Infant Development. San Antonio, TX, The Psychological Corporation, 1969
3. Campbell SK, Siegel E, Parr CA, Ramey CT: Evidence for the need to renorm the Bayley Scales of Infant Development based on the performance of a population-based sample of 12-month-old infants. Top Early Child Spec Educ 6:83–96, 1986
4. Nellis L, Gridley BE: Review of the Bayley Scales of Infant Development, 2nd ed. J Sch Psychol 32:201–209, 1994
5. Bayley N: Training Manual for the Bayley Scales of Infant Development, 2nd ed. San Antonio, TX, The Psychological Corporation, 1993
6. Ross G, Lawson K: Using the Bayley-II: Unresolved issues in assessing the development of prematurely born children. J Dev Behav Pediatr 18:109–111, 1997
7. Matula K, Gyurke JS, Aylward GP: Response to commentary: Bayley Scales-II. J Dev Behav Pediatr 18:112–113, 1997