Published in final edited form as:

Res Dev Disabil. 2014 January ; 35(1): . doi:10.1016/j.ridd.2013.10.017.

Instability of Delay Classification and Determination of Early Intervention Eligibility in the First Two Years of Life

Michele A. Lobo,

Physical Therapy Department, University of Delaware, Newark, DE, 19716

David A. Paul,

Thomas Jefferson University Medical College and Christiana Care Health Services, Christiana Neonatal Associates, MAP 1, Suite 217, Newark, DE 19718

Amy Mackley,

Christiana Care Health Services, Christiana Neonatal Associates, MAP 1, Suite 217, Newark, DE 19718

Jennifer Maher, and

Thomas Jefferson University Medical College and Christiana Care Health Services, Christiana Neonatal Associates, MAP 1, Suite 217, Newark, DE 19718

James Cole Galloway

Physical Therapy Department, University of Delaware, Newark, DE, 19716

Abstract

The purpose of this study was to determine the effectiveness of the Bayley Scales of Infant Development, Third Edition (Bayley-III) to track development and classify delays in low- and high-risk infants across the first two years of life. We assessed cognitive, language, and motor development in 24 low-risk full-term and 30 high-risk preterm infants via seven assessments performed between 3 and 24 months corrected age. The Bayley-III resulted in highly unstable delay classifications, low sensitivities, and poor positive predictive values across time. The results highlight that early intervention professionals, researchers, and policy makers should: (1) emphasize clinical opinion and prevalence of risk factors rather than standardized assessment findings when classifying delays and determining eligibility for services, and (2) develop more effective developmental assessments for infants and young children.

Keywords

assessment; early intervention; eligibility; developmental delay; preterm; Bayley Scales of Infant Development; Third Edition

© 2013 Elsevier Ltd. All rights reserved.

Corresponding concerning this article should be addressed to: Michele A. Lobo, Physical Therapy Department, University of Delaware, Newark, DE 19716. malobo@udel.edu.

The authors would like to express their sincere gratitude to the families who participated in this study.

The authors declare that there are no conflicts of interest.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The purpose of this study was to determine the effectiveness of repeated measurements using a common standardized assessment tool to track development in low- and high-risk infants in the first two years of life. This type of testing mirrors the standard of practice in the U.S. for assessing development when a potential delay has been noted or when high-risk infants, such as those born preterm or with brain injuries, receive developmental monitoring (Disabilities, 2006; Romanczyk et al., 2005). Each state has its own criteria for what defines “delay” and eligibility for early intervention services hinges on these criteria. In some cases, determination of eligibility is clear. For instance, if an infant is born with a diagnosis, such as Down syndrome, associated with future delays, the infant would qualify for services in most states. In other cases where outcomes are more heterogeneous so there is risk but not a certainty of future delay, it is more difficult to determine if an infant or child should receive services (Aarnoudse-Moens, Weisglas-Kuperus, van Goudoever, & Oosterlaan, 2009). This is the case for the growing population of infants born preterm.

Infants born preterm are more likely to have future cognitive, social, communication, and motor delays (Jongbloed-Pereboom, Janssen, Steenbergen, & Nijhuis-van der Sanden, 2012; Taylor et al., 2011). Therefore, their development is typically closely monitored in the first years of life through follow-up clinics and home visits provided at regular intervals, such as every 3–6 months (Janssen et al., 2011). The standard of practice at these visits is to assess development using a standardized assessment tool, such as the Bayley Scales of Infant Development, Third Edition (Bayley-III). The Bayley-III is a commonly selected tool for classification of developmental delay because it assesses development across multiple domains, is normed on young populations (1–42 months), has a long history of use in clinical and research settings, and is often considered a “gold standard” for early developmental assessment (Bayley, 2006b; Vincer et al., 2005).

The Bayley-III is a useful tool for classification of early developmental delay because it allows professionals to standardize an individual’s score to quantify degree of delay. This is important because if a child does not have a diagnosis clearly associated with future delays, early intervention eligibility is typically determined by his/her percentage of delay or the number of standard deviations his/her score lies below the mean. Therefore, the policy creates a demand for professionals to quantify developmental delay using specific measures. For all of these reasons, the Bayley-III was selected as the standardized assessment tool to track development in this study. Furthermore, the criterion for classification of delay was selected to be developmental performance more than 1.5 standard deviations below the mean because this is a common criterion in many states, such as New Mexico and Pennsylvania (New Mexico Administrative Code <http://nmhealth.org/ddsd/nmfit/Documents/documents/7.30.8%20NMAC%2006-29-12%20FINAL.pdf>; The Pennsylvania Code <http://www.pacode.com/secure/data/055/chapter4226/s4226.22.html>).

The two-year follow-up schedule, assessment tool, and delay criterion used in this study modeled the standard of practice and allowed us to test whether the Bayley-III is a valid and reliable means for the critical determination of a young child’s early intervention eligibility. Although the newest version of the Bayley is one of the assessments most commonly used in practice and research, this is the first study to our knowledge assessing development longitudinally with the Bayley-III in a cohort of low- and high-risk infants (Johnson & Marlow, 2006). Previous research has identified problems when using other early norm-referenced assessment tools, including the second version of the Bayley, with young populations at risk for delays. Some of these problems include low sensitivity, poor positive predictive value, large discrepancies in scores between two proximal testing points, and instability of delay classification across time (Hess, Papas, & Black, 2004; Horner, 1988; McGrath, Wypij, Rappaport, Newburger, & Bellinger, 2004; Vohr et al., 2012). Similar longitudinal studies have not been performed with the Bayley-III. Therefore, questions

remain about the validity, reliability, and utility of standardized assessment tools, and specifically the commonly used Bayley-III, for young populations.

Methods

Participants

Fifty-four infants and families participated in this study. Parents of all infants provided informed written consent. Twenty-four participants were born full-term without any diagnosed delays or medical risks and were recruited from the community. Thirty participants were born preterm and were recruited from a regional Level Three neonatal intensive care unit (Christiana Care Health Services, Newark, DE). All full-term infants were born between 37 and 41 weeks of gestation and all preterm infants were born between 24 and 30 weeks of gestation. The full-term group was free of diagnoses throughout the two-year study with the exceptions of asthma ($n=1$), allergies ($n=2$), and laryngomalacia ($n=1$). Diagnoses present in the preterm group throughout the study were periventricular leukomalacia ($n=6$), intraventricular hemorrhage ($n=8$), asthma ($n=3$), gastroesophageal reflux disease ($n=6$), bronchopulmonary dysplasia ($n=6$), retinopathy of prematurity ($n=2$), developmental delay ($n=3$), hypertrophic cardiomyopathy ($n=2$), reactive airway disease ($n=2$), hypertonia ($n=1$), hydrocephalus ($n=1$), chronic lung disease ($n=1$), eczema ($n=1$), spastic diplegic cerebral palsy ($n=1$), and hemiplegic cerebral palsy ($n=1$). Please see Table 1 for detailed demographic information about the groups.

Procedures

This study was a repeated measures assessment of development from three months through two years corrected age using the Bayley-III. The Bayley-III was administered in participants' homes when participants were alert and in a positive behavioral state. The cognitive, receptive language, expressive language, fine motor, and gross motor scales were administered at 3, 4, 6, 9, 12, 18, and 24 months of age. Ages were corrected for participants born preterm. The data reported are a subset from a larger study assessing early learning abilities in infants born at risk for future delays. Therefore, the early visits were more frequent. Assessments were performed by a licensed pediatric physical therapist with expertise in early intervention and a doctoral degree in development. The assessments were video recorded for future scoring by two trained pediatric professionals. Both scorers had masters degrees and knowledge about child development, were not present for home assessments, and were blind to participants' birth and medical histories. They maintained >95% intra- and inter-rater reliabilities across 20% of the scored visits (inter-rater reliability 96.2%; intra-rater reliabilities 97.1 and 96.6%). Administration and scoring training incorporated the Bayley-III Fundamental Administration DVD, in depth review of the Bayley-III Administration Manual, and practice testing/scoring.

Measures

The Bayley-III is comprised of subscales providing scores representing cognitive, language (expressive and receptive), motor (fine and gross), social-emotional, and adaptive behavioral development. The cognitive, language, and motor subscales are scored based on direct observations of child behavior, while the remaining subscales involve caregiver questionnaires (Bayley, 2006b). The present study examines the stability of only the subscales involving direct observations of child behavior.

The reliability and validity of the Bayley-III have been shown to be high. For instance, it has an average internal consistency of greater than .85 for all subscales, test-retest correlations for the subscales across ages ranges from .67 to .94, and its findings correlate highly with

findings from other standardized measures like the Wechsler Preschool and Primary Scale of Intelligence and the Preschool Language Scale (Bayley, 2006a).

Bayley-III items were each scored according to the Bayley-III Administration Manual. A participant's basal level was identified when ones were achieved for the first three items for an age group, meaning these behaviors were observed. A participant's ceiling level was identified when five consecutive zeros were scored for items, meaning these behaviors were not observed or were not performed correctly. Between the basal and ceiling levels, participants received ones for behaviors observed and performed correctly and 0's for behaviors not observed or performed incorrectly. For each subscale, a total raw score was calculated by summing the item scores. These raw scores were converted to normalized scaled scores using the Bayley III Administration Manual tables. The scaled scores range from one to 19, with a value of 10 representing the mean, lower scores representing poorer performance, and a quantity of three representing one standard deviation. Delay on a subscale was defined as a scaled score more than 1.5 standard deviations below the mean (< 5.5) because this is a common cut-off for early intervention eligibility in many states within the U.S.

Analyses

We used the Bayley-III delay classifications (delayed or not delayed) at each assessment age to first determine the stability of classification across time. If the classification was the same at every assessment age, the participant was categorized as having a Stable Classification. If there was one switch among classifications across time, the participant was categorized as having a Relatively Stable Classification. If there were multiple fluctuations with more than one switch among classifications across time, the participant was categorized as having an Unstable Classification (Janssen et al., 2011). Analysis of variance with planned comparison testing was performed to compare the distributions among these categories. Significance was set at $.05$. Effect sizes for significant findings are reported in terms of Pearson's r , with 0.10, 0.30, and 0.50 reflecting small, medium, and large effects respectively.

We also compared the Bayley-III delay classifications (delayed or not delayed) at each assessment age to the Bayley-III delay classification at the final assessment age to determine how well the Bayley-III identified and predicted future delay. For each subscale, we calculated: (1) *sensitivity*, or the proportion of children with delays at the final assessment who were identified as delayed at earlier assessments (the number of true positives divided by the number of true positives plus the number of false negatives), (2) *specificity*, or the proportion of children with typical development at the final assessment who were identified as having typical development at earlier assessments (the number of true negatives divided by the number of true negatives plus the number of false positives), (3) *positive predictive value*, or the proportion of children with delays at each assessment age who were also delayed at the final assessment (PPV; the number of true positives divided by the number of true positives plus the number of false positives), and (4) *negative predictive value*, or the proportion of children with typical development at each assessment age who were also typically developing at the final assessment (NPV; the number of true negatives divided by the number of true negatives plus the number of false negatives).

Results

Between 4 and 22% of the participants demonstrated delays in the five Bayley-III subscale domains. Table 2 shows the distribution of these delays.

Figure 1 depicts the stability of delay classifications for participants for each subscale. On average across the subscales, the classifications were Stable $44.1 \pm 8.6\%$ of the time (range

16.7–64.8%), Relatively Stable $13.0 \pm 3.4\%$ of the time (range 1.9–22.2%), and Unstable $43.0 \pm 6.7\%$ of the time (range 27.8–66.7%; Figure 1A). The percentage of participants in each category was unequal, $F(2, 12) = 7.162, p = .009, r = .74$. The percentage of participants with a Relatively Stable delay classification was less than the percentage with Stable ($p = .05, r = .76$) or Unstable classifications ($p = .002, r = .81$). The percentages of participants with Stable and Unstable classifications were not different ($p = 1.0$). The highest stability was observed for the fine motor subscale, potentially because most participants appeared to have typical development in this domain and the Bayley-III has been shown to quite successfully identify infants with typical development (Lobo & Galloway, 2013). Only four percent of the participants were delayed in the fine motor domain at 24 months.

One might expect a reliable and valid assessment to produce Stable classifications and at times Relatively Stable classifications, for instance when an infant has typical early development but due to environmental factors develops a delay. On the other hand, multiple fluctuations in delay classification are not generally expected. It is not expected, for instance, a child will be delayed at three months, not delayed at four months, and delayed again at five months. To compare expected outcomes with unexpected outcomes, we summed the percentage of participants with Stable or Relatively Stable classifications (expected patterns of outcomes for a reliable and valid classification system). We compared this with the percentage of participants with Unstable classifications (unexpected outcome) and found no difference, $t(4) = -1.042, p = .36$. This means that the assessment classified infants' development inconsistently with multiple shifts in classification just as often as it did consistently or in a meaningful way with one or fewer shifts in classification. For all subtests, the number of participants with Unstable delay classifications was substantial (Figure 1B).

Figure 2 shows the sensitivity, specificity, positive predictive value, and negative predictive value for each subtest across time. Note that across subtests the specificity and negative predictive value is generally high while the sensitivity and positive predictive value is generally low. This suggests the assessment is able to detect infants with typical development but is poor at identifying atypical development.

Discussion

The results of this study highlight the limitations of relying primarily on standardized assessment tools for identification of early developmental delays (McGrath et al., 2004; Vohr et al., 2012). In this group of low- and high-risk infants, all subscales of the Bayley-III resulted in highly unstable delay classifications, low sensitivities, and poor positive predictive values across time. Although the sample size for this study was modest and the number of infants who demonstrated fine motor delays was small, the results showed that it was common for an infant's delay classification to switch more than once throughout the first two years of life.

These findings suggest the practice of emphasizing standardized assessment scores does not accurately and reliably identify early developmental delays (Vohr et al., 2012). This poses critical challenges for early intervention professionals who need to determine eligibility for services and for those researching early delays. A common belief noted in the literature and practice is that early assessment should be an ongoing process involving multiple assessment points (Gartland & Strosnider, 2007; Noritz, Murphy, & Panel, 2013; Romanczyk et al., 2005). However, how does one make sense of multiple assessment points when they result in different classifications of delay? A change in delay might be expected if a child without biological risk is born into a high-risk environment and demonstrates a transition from the classification of "not delayed" to "delayed". It might also be expected if

a child is born with biological risks but receives effective early intervention and demonstrates a transition from “delayed” to “not delayed”. Yet, multiple fluctuations in classification like those observed in this study are more challenging to understand. Does a negative finding really mean the infant is not delayed and no services are needed? Does a positive finding really mean the infant is delayed and services should be provided? Should someone be classified as delayed if s/he meets the delay criteria once, or twice, or at specific ages? Multiple fluctuations make it almost impossible to make sense of repeated assessments.

We propose these results highlight the importance of educating professionals and policy makers to focus on eligibility criteria that deemphasize traditional norm-referenced standardized testing and emphasize other components of a comprehensive assessment, particularly clinical opinion and the presence of risk factors (Gartland & Strosnider, 2007; Romanczyk et al., 2005). Informed clinical opinion is already listed as a qualifier for services for infants and children in over half of the states in the U.S. This means that a professional can recommend services based on his/her global assessment. Professionals gain a more complete picture of the infant or child by incorporating observations across contexts, reports from parents, teachers, and other caregivers, and multiple assessment methods (Carey & Long, 2012; Gartland & Strosnider, 2007). It is important that early intervention professionals and policy makers understand that a global picture of the child provides a better determination of the child’s abilities than do the results of one standardized assessment. Furthermore, when the standardized test results conflict with the observations, reports, and other global findings, early intervention eligibility and other critical decisions should be determined based on the global assessment and the clinical opinion of the early intervention professionals (Msall, Tremont, & Ottenbacher, 2001; Romanczyk et al., 2005).

The presence of risk factors for future delay should also be emphasized when determining early intervention eligibility, particularly in the first months and year of life (Romanczyk et al., 2005). Figure 3 lists some early prenatal, perinatal, neonatal, postnatal, social, and behavioral risk factors that relate to future delays, impairments, and diagnoses. The number and severity of risk factors an infant has can help guide the level of service delivery provided. For instance, if an infant is born with a brain injury, s/he might start with intervention and close monitoring services immediately after hospital discharge (Drougia et al., 2007). If an infant is born preterm at 34 weeks gestation without other risk factors, close monitoring of that infant at 3-6-month intervals might start after discharge without direct intervention (Petrini et al., 2009). However if that infant also has bronchopulmonary dysplasia or socioeconomic risks, intervention and close monitoring services might be started immediately after hospital discharge (Lee, 2011; Natarajan et al., 2012).

The findings of this study also highlight the need for designing better assessment tools for early intervention practice and research. The literature suggests the most valid and reliable tools focus on caregiver input, exploratory play, and spontaneous function, participation, and social interaction with peers across contexts (Lobo & Galloway, 2013; Msall et al., 2001). For instance, observation of spontaneous play and performance using play-based assessments can be more sensitive for identifying early delays than standardized assessments like the Bayley (Heathcock, Bhat, Lobo, & Galloway, 2004; Kodric, Sustersic, & Paro-Panjan, 2010; Lobo & Galloway, 2013). This is in contrast to most norm-referenced standardized assessments, which involve the performance of discrete tasks presented in a predetermined manner in isolated environments by adult examiners (Bayley, 2006b). For example, software could be created for use with novel motion tracking technology like the Kinect system to assess the amount, type, and variability of movement, play, and socializing in natural circumstances (Dutta, 2012). In addition, early intervention providers could choose more naturalistic, play-based assessments such as the Individual Growth and

Development Indicators for Infants and Toddlers or the Infant Preschool Play Assessment Scale (Flagler, 1996; Walker, Carta, Greenwood, & Buzhardt, 2008). In the meantime, it is imperative that early intervention policy makers and institutions funding early intervention research recognize the limitations of the current standardized assessment tools and the value of clinical judgment that incorporates multiple sources of data, including observations across contexts, parent reports, and play-based assessments that are gaining empirical support (Lobo & Galloway, 2013; Simard, Luu, & Gosselin, 2012).

Acknowledgments

This research was supported by the National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health & Human Development, 1R01HD051748 (Motor learning and coordination in high-risk infants. Principal investigator: Cole Galloway, PT, PhD, August 2007-June 2012).

References

- Aarnoudse-Moens CSH, Weisglas-Kuperus N, van Goudoever JB, Oosterlaan J. Meta-Analysis of Neurobehavioral Outcomes in Very Preterm and/or Very Low Birth Weight Children. *Pediatrics*. 2009; 124(2):717–728.10.1542/peds.2008-2816 [PubMed: 19651588]
- Bayley, N. Technical Manual. 3. San Antonio, TX: Pearson; 2006a. Bayley Scales of Infant and Toddler Development.
- Bayley, N. Bayley Scales of Infant Development. 3. San Antonio, TX: The Psychological Corporation; 2006b.
- Carey H, Long T. The Pediatric Physical Therapist's Role in Promoting and Measuring Participation in Children With Disabilities. *Pediatric Physical Therapy*. 2012; 24(2):163–170.10.1097/PEP.0b013e31824c8ea2 [PubMed: 22466385]
- Disabilities, CoCW. Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. *Pediatrics*. 2006; 118(1):405–420. [PubMed: 16818591]
- Drougia A, Giapros V, Kralli SN, Theochari SP, Nikaki A, Tzoufi M, Andronikou S. Incidence and risk factors for cerebral palsy in infants with perinatal problems: A 15-year review. *Early Human Development*. 2007; 83(8):541–547. [PubMed: 17188824]
- Dutta T. Evaluation of the Kinect (TM) sensor for 3-D kinematic measurement in the workplace. *Applied Ergonomics*. 2012; 43(4):645–649.10.1016/j.apergo.2011.09.011 [PubMed: 22018839]
- Flagler, SL. Multidimensional assessment of young children through play. Lewisville, TX: Kaplan Press; 1996.
- Gartland D, Strosnider R. Learning disabilities and young children: Identification and intervention. *Learning Disability Quarterly*. 2007; 30(1):63–72.
- Heathcock JC, Bhat AN, Lobo MA, Galloway JC. The performance of infants born preterm and full-term in the mobile paradigm: Learning and memory. *Physical Therapy*. 2004; 84(9):808–821. [PubMed: 15330694]
- Hess CR, Papas MA, Black MM. Use of the bayley infant neurodevelopmental screener with an environmental risk group. *Journal of Pediatric Psychology*. 2004; 29(5):321–330.10.1093/jpepsy/jsh036 [PubMed: 15187171]
- Horner TM. Single versus repeated assessments of infant abilities using the Bayley Scales of Infant Development. *Infant Mental Health Journal*. 1988; 9(3):209–217.10.1002/1097-0355(198823)9:3<209::aid-imhj2280090304>3.0.co;2-e
- Janssen AJWM, Akkermans RP, Steiner K, de Haes OAM, oostendorp RAB, Kollee LAA, Nijhuis-van der Sanden MWG. Unstable longitudinal motor performance in preterm infants from 6 to 24 months on the Bayley Scales of Infant Development - second edition. *Research in Developmental Disabilities*. 2011; 32:1902–1909. [PubMed: 21530160]
- Johnson S, Marlow N. Developmental screen or developmental testing? *Early Human Development*. 2006; 82(3):173–183.10.1016/j.earlhumdev.2006.01.008 [PubMed: 16504424]

- Jongbloed-Pereboom M, Janssen A, Steenbergen B, Nijhuis-van der Sanden MWG. Motor learning and working memory in children born preterm: A systematic review. *Neuroscience and Biobehavioral Reviews*. 2012; 36(4):1314–1330.10.1016/j.neubiorev.2012.02.005 [PubMed: 22353425]
- Kodric J, Sustersic B, Paro-Panjan D. Assessment of general movements and 2.5 year developmental outcomes: Pilot results in a diverse preterm group. *European Journal of Paediatric Neurology*. 2010; 14(2):131–137.10.1016/j.ejpn.2009.04.012 [PubMed: 19540782]
- Lee JS. The Effects of Persistent Poverty on Children's Physical, Socio-emotional, and Learning Outcomes. *Child Indicators Research*. 2011; 4(4):725–747.10.1007/s12187-011-9120-8
- Lobo MA, Galloway JC. Assessment and stability of early learning abilities in preterm and full-term infants across the first two years of life. *Research in Developmental Disabilities*. 2013; 34:1721–1730. [PubMed: 23500166]
- McGrath E, Wypij D, Rappaport LA, Newburger JW, Bellinger DC. Prediction of IQ and achievement at age 8 years from neurodevelopmental status at age 1 year in children with D-transposition of the great arteries. *Pediatrics*. 2004; 114(5):E572–E576.10.1542/peds.2003-0983-L [PubMed: 15492354]
- Msall ME, Tremont MR, Ottenbacher KJ. Functional assessments of preschool children: Optimizing developmental and family supports in early intervention. *Infants and Young Children*. 2001; 14(1): 46–66.
- Natarajan G, Pappas A, Shankaran S, Kendrick DE, Das A, Higgins RD, Walsh MC. Outcomes of extremely low birth weight infants with bronchopulmonary dysplasia: Impact of the physiologic definition. *Early Human Development*. 2012; 88(7):509–515.10.1016/j.earlhumdev.2011.12.013 [PubMed: 22236557]
- Noritz GH, Murphy N, Panel NSE. Motor delays: Early identification and evaluation. *Pediatrics*. 2013
- Petrini JR, Dias T, McCormick MC, Massolo ML, Green NS, Escobar GJ. Increased Risk of Adverse Neurological Development for Late Preterm Infants. *Journal of Pediatrics*. 2009; 154(2):169–176.10.1016/j.jpeds.2008.08.020 [PubMed: 19081113]
- Romanczyk RG, Gillis JM, Noyes-Grosser DM, Holland PP, Holland CL, Lyons D. Clinical clues, developmental milestones, and early identification/assessment of children with disabilities: Practical applications and conceptual considerations. *Infants and Young Children*. 2005; 18(3): 212–221.
- Simard MN, Luu TM, Gosselin J. Concurrent Validity of Ages and Stages Questionnaires in Preterm Infants. *Pediatrics*. 2012; 130(1):E108–E114.10.1542/peds.2011-3532 [PubMed: 22689873]
- Taylor HG, Klein N, Anselmo MG, Minich N, Espy KA, Hack M. Learning problems in kindergarten students with extremely preterm birth. *Archives of pediatrics & adolescent medicine*. 2011; 165(9):819–825. [PubMed: 21893648]
- Vincer MJ, Cake H, Graven M, Dodds L, McHugh S, Fraboni T. A population-based study to determine the performance of the Cognitive Adaptive Test/Clinical Linguistic and Auditory Milestone Scale to predict the mental developmental index at 18 months on the Bayley Scales of Infant Development-II in very preterm infants. *Pediatrics*. 2005; 116(6):E864–E867.10.1542/peds.2005-0447 [PubMed: 16263973]
- Vohr BR, Stephens BE, Higgins RD, Bann CM, Hintz SR, Das A, Eunice Kennedy Shriver Natl I. Are Outcomes of Extremely Preterm Infants Improving? Impact of Bayley Assessment on Outcomes. *Journal of Pediatrics*. 2012; 161(2):222.10.1016/j.jpeds.2012.01.057 [PubMed: 22421261]
- Walker D, Carta JJ, Greenwood CR, Buzhardt JF. The Use of Individual Growth and Developmental Indicators for Progress Monitoring and Intervention Decision Making in Early Education. *Exceptionality*. 2008; 16(1):33–47.10.1080/09362830701796784

Highlights

- We assessed development longitudinally in low- and high-risk infants using the Bayley-III from 3 to 24 months.
- It was common for Bayley-III delay classifications to switch multiple times across the first two years.
- The Bayley-III was effective at identifying typical development but poor at identifying infants with delays.
- Clinical opinion and risk factors should be valued above standardized assessments for identifying early delays.

Figure 1.

Percentages of participants whose classification of delay was Stable (no change), Relatively Stable (one change in classification), or Unstable (more than one change in classification) across time (A) averaged across the five subscales of the Bayley-III and (B) for the individual subscales. Delay represented a scaled score more than 1.5 standard deviations below the mean. Note that the percentage of Unstable classifications did not differ from the percentage of Stable classifications, nor did it differ from the summed percentage of Stable plus Relatively Stable classifications. Also note that a large percentage of the participants had multiple changes in delay classification across all of the Bayley-III subtests. Error bars reflect standard error of the mean. There are no error bars in B because data represent totals of infants within each subscale, not averages.

Figure 2. Relationships between Bayley-III delay classifications from 3-18 months and the Bayley-III delay classification at 24 months. Sensitivities, specificities, positive predictive values (PPV), and negative predictive values (NPV) are shown for each of the subscales. Note that the subtests are generally strong at identifying typical development (high specificities and NPVs) but poor at identifying atypical development (low sensitivities and PPVs).

RISK FACTORS	
A) Prenatal Maternal Factors	
1) Cocaine Exposure ¹	
2) Fetal Alcohol Syndrome ²	
3) Lead Toxicity ³	
4) Mercury Toxicity ⁴	
5) Valproate Toxicity ⁵	
6) Severe Stress ⁶	
B) Perinatal Factors	
1) Hypoxic Ischemic Encephalopathy ⁷	
2) Prematurity: <28 Weeks ⁸	
3) Prematurity: 28-33 Weeks ⁹	
4) Prematurity: 34-36 Weeks ¹⁰	
5) Low Birth Weight (<1500g) ¹¹	
6) ELBW (<1000g) ¹²	
7) Multiple Gestation ¹³	
C) Neonatal Factors	
1) Intracranial Hemorrhage ¹⁴	
2) Meningitis ¹⁵	
3) Congenital Infection: Herpes simplex virus ¹⁶	
4) Congenital Infection: Cytomegalovirus ¹⁷	
5) Congenital Infection: Toxoplasmosis ¹⁸	
6) Feeding Problems ¹⁹	
7) Bronchopulmonary dysplasia/ Chronic Lung Disease ²⁰	
D) Postnatal Factors	
1) Seizures ²¹	
2) Meningitis ²²	
3) Malnutrition ²³	
4) Lead Exposure ²⁴	
E) Factors in Social History	
1) History of abuse/neglect ²⁵	
2) Limited social or financial support ²⁶	
3) Maternal Depression ²⁷	
4) Teenage Parent ²⁸	
5) Low Parent Education ²⁹	
F) Behavioral Factors	
1) Delayed Sitting ³⁰	
2) Delayed Walking ³¹	
3) Delayed Speaking ³²	
4) Regression in Speech Development ³³	
5) Toe-Walking ³⁴	

Figure 3.

Common prenatal, perinatal, neonatal, postnatal, social, and behavioral risk associated with future delays, impairments, and diagnoses.

¹(Bauer et al., 2005; Bennett, Bendersky, & Lewis, 2002; Davis et al., 1992; Heffelfinger, Craft, & Shyken, 1997; T. P. Levine et al., 2008; Ostrea, Ostrea, & Simpson, 1997; L. T. Singer et al., 2002)

²(Kalberg et al., 2006; Kvigne et al., 2004; Stromland, 2004)

³(Hu et al., 2006)

⁴(Axelrad, Bellinger, Ryan, & Woodruff, 2007; Grandjean et al., 1999; Stromland, 2004)

⁵(Cohen et al., 2011; Dean et al., 2002; Eriksson et al., 2005; Rasalam et al., 2005; Thomas et al., 2008)

⁶(Kinney, Munir, Crowley, & Miller, 2008; Laplante, Zelazo, Brunet, & King, 2007)

⁷(Martinez-Biarge, Diez-Sebastian, Rutherford, & Cowan, 2010; Shankaran et al., 2005; Vannucci, 2000)

⁸(Hack & Fanaroff, 2000; Rugolo, 2005)

- ⁹(Ancel et al., 2006; Bracewell & Marlow, 2002)
- ¹⁰(McIntire & Leveno, 2008; Petrini et al., 2009)
- ¹¹(Hack et al., 2004; Hendson et al., 2011; Horbar et al., 2012; Procianoy, Koch, & Silveira, 2009)
- ¹²(Gianni et al., 2007; Hintz et al., 2005; Laptook, O'Shea, Shankaran, Bhaskar, & Network, 2005)
- ¹³(Baergen, 2011; Hillemeier, Morgan, Farkas, & Maczuga, 2011; Scher et al., 2002)
- ¹⁴(De Vries, Van Haastert, Rademaker, Koopman, & Groenedaal, 2004; O'Shea, Klinepeter, Goldstein, & Dillard, 1997)
- ¹⁵(Edmond et al., 2010; Klinger, Chin, Beyene, & Perlman, 2000; Libster et al., 2012)
- ¹⁶(Kimberlin, 2007)
- ¹⁷(Bale, Miner, & Petheram, 2002; Coats et al., 2000; Dollard, Grosse, & Ross, 2007)
- ¹⁸(Roizen et al., 1995)
- ¹⁹(Chatoor et al., 2004)
- ²⁰(Doyle & Anderson, 2009; Kinsella, Greenough, & Abman, 2006; Short et al., 2003; L. Singer, Yamashita, Lilien, Collin, & Baley, 1997)
- ²¹(Russ, Larson, & Halfon, 2012; Tekgul et al., 2006)
- ²²(Singhi, Bansal, Geeta, & Singhi, 2007)
- ²³(Park et al., 2011)
- ²⁴(Carlisle, Dowling, Siegel, & Alexeeff, 2009)
- ²⁵(Hanson, Jawad, Ryan, & Silver, 2011)
- ²⁶(Bradley & Corwyn, 2002; Hillemeier et al., 2011)
- ²⁷(Cornish et al., 2005)
- ²⁸(Clements, Barfield, Kotelchuck, & Wilber, 2008; Keown, Woodward, & Field, 2001; J. A. Levine, Pollack, & Comfort, 2001; Sommer et al., 2000)
- ²⁹(Hillemeier et al., 2011)
- ³⁰(Harbourne, Willett, Kyvelidou, Deffeyes, & Stergiou, 2010; Jones, Morgan, Shelton, & Thorogood, 2007)
- ³¹(Jahnsen, Villien, Egeland, Stanghelle, & Holm, 2004; Murray, Jones, Kuh, & Richards, 2007; Palmer, 2004)
- ³²(Chuang et al., 2011; Mandell, Novak, & Zubritsky, 2005)
- ³³(Ozonoff, Williams, & Landa, 2005; Pickles et al., 2009; Shinnar et al., 2001)
- ³⁴(Barrow, Jaworski, & Accardo, 2011; Engstrom, Hoof, & Tedroff, 2012; Rodda & Graham, 2001)

Table 1

Demographic data

Characteristic	Low Risk Participants	High Risk Participants
Number of males	14 (58%)	10 (33%)
Number of females	10 (42%)	20 (67%)
Number of Caucasians	17 (71%)	12 (40%)
Number of Blacks	5 (21%)	13 (43%)
Number of Asians	2 (8%)	5 (17%)
Number of Hispanics	0 (0%)	4 (13%)
Age at birth (weeks)	39.4 ± .2	26.7 ± .3
Number born very preterm (<32 weeks)	0 (0%)	9 (30%)
Number born extremely preterm (<28 weeks)	0 (0%)	21 (70%)
Weight at birth (grams)	3366.7 ± 135.8	921.3 ± 47.5
Number with very low birth weight (<1500 grams)	0	10 (33%)
Number with extremely low birth weight (<1000 grams)	0	20 (67%)

Table 2

Prevalence of delays (more than 1.5 standard deviations below the mean) on the Bayley-III and provision of early intervention services across domains at 24 months.

Subscale	Low Risk Participants	High Risk Participants	Total Participants
Cognitive Delay	0 (0%)	9 (30%)	9 (17%)
Special Education Services	0 (0%)	2 (7%)	2 (4%)
Receptive Language Delay	3 (13%)	9 (30%)	12 (22%)
Expressive Language Delay	2 (8%)	5 (17%)	7 (13%)
Speech Therapy Services	0 (0%)	5 (17%)	5 (9%)
Fine Motor Delay	0 (0%)	2 (7%)	2 (4%)
Occupational Therapy Services	0 (0%)	5 (17%)	5 (9%)
Gross Motor Delay	3 (13%)	6 (20%)	9 (17%)
Physical Therapy Services	0 (0%)	7 (23%)	7 (13%)