The Bayley Scales: Clarification for Clinicians and Researchers

Glen P. Aylward, PhD, ABPP, and Jiajun Zhu, PhD

Introduction

The Bayley Scales have been the developmental reference standard since 1969 and are used in clinical assessment and research to measure developmental delays, to measure outcomes of specific risk groups, and measure the effects of environmental and medical interventions. Because there are now four editions of the Bayley Scales with some discrepancies in reported scores, this has caused confusion in pediatrics regarding which edition is most accurate.

Inconsistency in scores between editions of the Bayley Scales have been attributed to the Flynn effect, over- and underestimation of normative data, differing basal and ceiling criteria, and changes in formats. For instance, the Bayley Scales of Infant Development™ (2nd ed.; Bayley™–II; Bayley, 1993) Mental Developmental Index (MDI) and Psychomotor Developmental Index (PDI) scores were lower than those same scores for the Bayley Scales of Infant Development (BSID; Bayley, 1969), whereas the Bayley Scales of Infant and Toddler Development (3rd ed.; Bayley–III; Bayley, 2006) scores were higher than the corresponding Bayley–II scores, prompting much effort to reconcile the Bayley–III with the Bayley–II scores. Consequently, the Bayley–IIIwas criticized because of suspected "inflated" norms and diagnostic inaccuracy, although multiple other factors could explain these higher scores (Anderson & Burnett, 2017; Anderson et al., 2010; Aylward & Aylward, 2011). These discrepancies caused difficulty in day-to-day clinical diagnostics, longitudinal programs that involve high-risk populations (e.g., premature children), or studies that assess outcomes following medical or environmental interventions. Nonetheless, the Bayley–III continued to be the reference standard for developmental assessment. Fueling the debate, the Bayley–III and new Bayley Scales of Infant and Toddler Development (4th ed.; Bayley-4; Bayley & Aylward, 2019) have comparable scores, despite significant changes.

Possible Reasons for Discrepancies Between the BSID, Bayley–II, and Bayley–III Scores

Comparing the BSID to the Bayley–II when administered in a counterbalanced order to 200 children revealed a decrease of 11.8 and 10.1 points for the MDI and PDI, respectively. Differences at the extremes were found to be as high as 30 points in some studies (Lennon, Gardner, Karmel, & Flory, 2008). One explanation for the decrease in scores is the Flynn effect (Flynn, 1999), although it is questionable whether this phenomenon and its causes are as applicable to infants and toddlers as to older populations (Trahan, Stuebing, Fletcher, & Hiscock, 2014). There is also new evidence that

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

the Flynn effect has more recently weakened or even reversed (Bratsberg & Rogeburg, 2018; Dutton, van der Linden, & Lynn, 2016) and that intelligence quotient (IQ) scores received a transient boost from environmental factors that masked a very subtle decline (Dutton et al., 2016). The most likely cause to explain the Flynn effect in infants and toddlers (out of the eight causes most frequently noted) is improvements in nutrition (Trahan et al., 2014; Lynn, 2009). Moreover, a review of the "infant Flynn effect" revealed the mean increase in developmental quotient (DQ) points over a decade was 3.4 points for the Cognitive Scale and 2.49 points for the Motor Scale (Lynn, 2009). The increase in DQ of the Griffiths Mental Development Scales-Revised (Griffiths & Huntley, 1996) was 2.45 (Lynn, 2009). Of particular note is an Australian sample that had larger gains—5.8 DQ points per decade for the Bayley–II in 1997 (Trahan et al., 2014), an issue that was also reported recently (Anderson et al., 2010). It also appears that different estimates of the Flynn effect are found on different tests.

When the Bayley–II and Bayley–III scores were compared in a sample of 102 children (Bayley, 2006), the Bayley–III Cognitive composite score equivalent was 7.1 points higher than the MDI and the Motor composite score was 8.4 points higher than the PDI; a surprising finding. Speculation regarding the cause of possible norm inflation in the Bayley–III included the switch from a two-subtest structure (i.e., Mental Developmental Scale and Psychomotor Developmental Scale) to a five-subtest structure (i.e., Cognitive, Receptive Communication, Expressive Communication, Fine Motor, and Gross Motor), the Flynn effect (Flynn, 1999), changes in the demographics of the population from 1988 to the 2000 census (including parent education level), differences in norming methodology (classical observed data approach vs. inferential or continuous norming approach), and the result of seeding the normative sample with 10% of at-risk children to avoid truncating norms (Pearson, 2008). The seeding of the normative sample with at-risk children did not occur during the development of the BSID or Bayley–II.

Introduction of the Bayley-4

Evaluating the reported discrepancies between the BSID, Bayley–II, and Bayley–III scores is made more complex with the recent counterbalanced comparison of the Bayley–III and the Bayley-4 (Bayley & Aylward, 2019) involving 184 children (mean age of 19.2 months and mean test interval of 10.3 days). As shown in Table 1, the Bayley-4 Cognitive, Language, and Motor Scales are comparable to the Bayley–III; Bayley-4 scores are 0.1 scaled score points and 0.5 and 1.3 standard score points lower than the Bayley–III, respectively. This similarity in norms across the Bayley–III and Bayley-4 continued the concerns regarding the normative changes originally raised with the Bayley–III. Whenever changes are made to an instrument, normative changes may result and clinicians need to be aware of how changes impact scores and interpretation. However, given the concerns with the Bayley–III norm shifts, clinicians were left without guidance on how to interpret the differences in scores, how to determine which test is most appropriate for a specific child, and how to determine if scores obtained from different Bayley Scales versions are interchangeable.

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

Table 1. Comparison of Bayley–III and Bayley-4 Cognitive, Language, and Motor Scores

Score	Bayley-4	Bayley–III	Difference	Corrected *r*
Cognitive[a]	10.4[a]	10.5[a]	-0.1	.70
Cognitive[b]	102.0[b]	102.5[b]	-0.5	.70
Language[b]	101.9[b]	102.4[b]	-0.5	.72
Motor[b]	101.0[b]	102.3[b]	-1.3	.75

[a]Scaled score (*M* = 10, *SD* = 3)
[b]Standard score (*M*= 100, *SD* = 15)

Regarding the question of which edition of the Bayley Scales is most discrepant, it is noteworthy that the Bayley–II may be overlooked as a possibility. Instead, there has been much effort exerted to try to bring the Bayley–III scores in line with the Bayley–II, implicating that the Bayley–II scores are more accurate. This impression drew further support from a study (Anderson et al., 2010) where an extremely preterm/extremely low birth weight (EPT/ELBW) Australian sample obtained a Bayley–III Cognitive composite with a mean of 96.9 and the matched control group of 108.9. However, close perusal of the extant literature raises some important, potentially conflicting conclusions.

An alternative explanation to the Flynn effect lowering the Bayley–II scores involves basal and ceiling rules and administered item sets (Lennon et al., 2008). The BSID was designed as a modified power test, containing items in a specific sequence that extend across several developmental domains. The basal/ceiling criterion was 10 consecutive items in a row passed/failed for the Mental Scale and six for the Motor Scale. Because items from different domains were interspersed in the item set, a child could fail some items but then pass others, thereby increasing his or her scores because the ceiling of 10 consecutive failed items was not met. In contrast, the Bayley–II had age-based item sets and selecting which item set to begin with was often at the discretion of the examiner. If the child started at a lower item set, he or she could fail some items that otherwise would have been automatically credited as passed if a higher item set had been selected. Paradoxically, correction for prematurity could work in an opposite direction (Gauthier, Bauer, Messinger, & Closius, 1999) because of this quirk. For example, the basal criterion for the Mental Scale was to pass five or more items in the item set and the ceiling was to fail three or more items in the item set; the basal and ceiling for the Motor Scale were four passes and two fails, respectively. No credit was given for passing items above the item set, thereby truncating scores (Lennon et al., 2008). In contrast, on the Bayley–III the child would need to pass three consecutive items at the age-based start point to establish a basal and fail five consecutive items to establish a ceiling. This again provided an opportunity for a higher score, similar to the BSID. The Bayley-4 has similar basal and ceiling criterion (i.e., three consecutive item scores of 2 for the basal and five consecutive scores of 0 for the ceiling). In fact, Lennon et al. (2008) stated, “lower scores on the Bayley–II may in fact have been due to the item set problem, since the restricted range of items presented during the administration of the Bayley–II was less likely to yield scores reflecting the true range of the child’s abilities. The extended range of functioning allowed by the original Bayley and the Bayley–III may contribute to higher and presumably more accurate scores on both” (p. 44).

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

Synnes et al. (2010) investigated changes in outcome scores in four age cohorts: 1983–1987, 1988–1992, 1993–1997, and 1998–2003. The BSID was used in the first two cohorts and the Bayley–II in the latter two cohorts. The MDI decreased from 93 to 85 whereas the PDI decreased from 89 to 78 over the four time periods (despite a lowered rate of cerebral palsy). The authors suggested "our results were affected by differences between BSID editions" (p. 992) as a possible reason for the decline, raising the question of whether these findings were due to a true worsening of outcome or measurement issues, particularly that the Bayley–II norms were too conservative.

Similarly, in a small study of 18-month-old children with EPT/ELBW who were evaluated with the Bayley–II and Bayley–III, the mean Bayley–II MDI was 89.4 whereas the Bayley–III Cognitive composite score equivalent was 96.5 (Robertson, Hendson, Biggs, & Acton, 2010). It was assumed that the Bayley–III scores were inaccurate, but once again, perhaps the Bayley–II scores were too conservative. Further questioning of the validity of Bayley–II scores is found in a report of children with ELBW assessed at age 20 months with the Bayley–II and the Short Forms of the Kaufman Assessment Battery for Children (KABC™) Mental Processing and Achievement Scales at age 8 years (Kaufman & Applegate, 1988). The predictive utility of low scores obtained on the Bayley–II was poor and the authors stated, "We are concerned…by reported high rates of cognitive impairments based on the use and presumptive validity of the Bayley–II MDI" (Hack et al., 2005, p. 333).

Also important is the finding in the literature that shows a variety of samples of typically developing children had mean Bayley–II scores that hovered around 95 (Connolly et al., 2006; McClain, Provost, & Crowe, 2000) and typically did not reach 100. There had also been some concern during data collection for the Bayley–II that recruitment methods (e.g., using published birth announcements) tended to systematically include high-end cases, thereby leading to the development of much harder or conservative norms. For example, Lennon et al. (2008) indicated that the Bayley–II normative sample "consisted of children who were increasingly higher functioning" and "contained very few low-scoring children" (p. 42).

Despite these concerns, various attempts were made to reconcile Bayley–II and Bayley–III scores, using the former as the reference standard. These attempts included combining the Bayley–III Cognitive and Language scores into a single composite (Moore, Johnson, Haider, Hennessy, & Marlow, 2012), developing algorithms (Moore et al., 2012) and conversion formulae (Lowe, Erickson, Schrader, & Duncan, 2012), applying DQ scores (Milne, McDonald, & Comino, 2012), changing the cut scores to 80 or 85 to indicate impairment (Johnson, Moore, & Marlow, 2014; Vohr et al., 2012),and an effort to renorm the Bayley–III or publish a new edition (Anderson & Burnett, 2017). In addition, some investigators suggested using cut scores based on local reference data (Spencer-Smith, Spittle, Lee, Doyle, & Anderson, 2015). In a recent editorial on the Bayley–III it was suggested that no further effort be expended to attempt to make scores more compatible between the two tests, underscoring the possibility that there may be some inflation of Bayley–III norms but also that the Bayley–II norms underestimate development (Aylward, 2013). This was also stated by Bos (2013), who wrote, "The question is whether the Bayley–II underestimates or the Bayley–III overestimates development. Both might be true" (p. 978).

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

An interim option mentioned by several investigators with respect to the Bayley–III was to compare the rates of impairment using several cut scores (e.g., 1 *SD* below the mean, 1.5 *SD* below the mean.) However, this approach is indirectly influenced by the assumption that the Bayley–II norms are most accurate. The other alternative is to publish a new edition, which has been done with the 2019 release of the Bayley-4.

Evidence for the Validity of the Bayley-4 Scores

Tests per se are not valid or invalid; rather, the test scores are either valid or invalid for making specific interpretations. Using test scores that are validated based on accumulating evidence provides a sound scientific basis for score interpretation. Three primary ways to achieve this goal are: (1) validate that the normative sample represents the current population accurately, (2) prove that the test scores are consistent with other measures tapping similar constructs, and (3) show that the test scores have clinical utility.

Representativeness of the Bayley-4 Normative Sample

The Bayley-4 normative sample is stratified according to the 2017 census data by age, sex, race/ethnicity, and parent education level. Trained recruiters and independent examiners identified children who met the specified inclusion criteria according to the selected demographic variables. Random case selections were used when multiple candidates met the same inclusion criteria and demographic requirements (Bayley & Aylward, 2019). Besides matching with the census data, the normative sample was also validated using average zip code income. The average zip code income of the Bayley-4 normative sample is $62,835, which is very close to the U.S. national average of $62,175 in 2017. To ensure the representation of full ability range, the Bayley-4 normative sample includes 21 cases diagnosed with Down syndrome (1.2%).

Logically, it is necessary to compare the two latest versions of the Bayley Scales; however, this comparison may not be convincing to some examiners in light of the criticisms of Bayley–III norms. In response, the Bayley–III norms were reviewed to assess what effects the inclusion of 10% of at-risk children had on the normative data. Table 2 shows that the inclusion of at-risk children results in very little change in the mean scores for the Cognitive, Language, and Motor Scales, which suggests the inclusion of at-risk children in the normative sample did not have a major impact on the assumed inflation of Bayley–III scores.

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

Table 2. Bayley–III Scores With and Without At-Risk Children in the Normative Sample

With at-risk	*N*	Mean	*SD*	Minimum	Maximum
Cognitive[a]	1700	9.9[a]	3.1	1	19
Language[b]	1700	100.4[b]	15.2	47	153
Motor[b]	1700	99.8[b]	15.1	46	154
Without at-risk	***N***	**Mean**	***SD***	**Minimum**	**Maximum**
Cognitive[a]	1534	10.1[a]	2.9	1	19
Language[b]	1534	101.0[b]	14.5	53	153
Motor[b]	1534	100.8[b]	14.3	46	154

[a] Scaled score (*M* = 10, *SD* = 3)

[b] Standard score (*M*= 100, *SD* = 15)

As previously mentioned, the Cognitive, Language, and Motor Scales scores of the Bayley–III and Bayley-4 were compared and there is very little difference between the scores. This high level of agreement may be facilitated to some degree by the similar formats (i.e., five-subtest structure), similar basal and ceiling rules, as well as several similar items (although there was also item overlap between the BSID and Bayley–II). Conversely, the Bayley-4 has a different scoring system, some different items, less redundancy in item content, uses caregiver report for a number of items, and normative data collection used the Bayley-4 on Q-global®. Although the possibility exists that the scores are similar for different reasons, the main point is that they are comparable.

Comparison of Bayley-4 to the WPPSI–IV and PDMS–2

As indicated in Table 3, the Bayley-4 scores are consistent with the scores on the Wechsler Preschool and Primary Scale of Intelligence (4th ed.; WPPSI®–IV; Wechsler, 2012) and the Peabody Developmental Motor Scales (2nd ed.; PDMS–2; Folio & Fewell, 2000).

WPPSI–IV

The Bayley-4 and WPPSI–IV were administered to 104 children as part of the development of the Bayley-4 (mean age of 36.6 months, mean test interval of 8.1 days). The mean Bayley-4 Cognitive scaled score was 9.7, which equates to a standard score of 98.5 and the mean WPPSI–IV FSIQ was 103.3. The WPPSI–IV Visual Spatial Index (VSI) was 100.2 compared to the aforementioned Cognitive standard score equivalent of 98.5. Lastly, the Bayley-4 Language standard score was 100.6, whereas the WPPSI–IV Verbal Comprehension Index (VCI) was 103.0. Using a statistically derived adjustment for the Flynn effect (Weiss, Gregoire, & Zhu, 2016) of .31 points per year for the FSIQ, .39 points per year for the VSI, and .13 points per year for the VCI, the differences between scores are 2.17, 2.73, and .91 points respectively, further reducing the difference between the Bayley-4 and WPPSI–IV scores.

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

PDMS–2

The Bayley-4 and PDMS–2 were administered to 100 children as part of the development of the Bayley-4 (mean age of age 18.6 months, mean test interval of 10.9 days). Both the Bayley-4 Motor standard score and the PDMS-2 Total Motor Quotient were 99.5. The Bayley-4 Fine Motor subtest had a scaled score of 9.9 which equates to a standard score of 99.5 and the PDMS–2 Fine Motor Quotient was 97.5. The Bayley-4 Gross Motor subtest scaled score was also 9.9 (equating to a standard score of 99.5), whereas the PDMS–2 Gross Motor Quotient was 101.1. Similar to the WPPSI–IV, the PDMS–2 scores were consistent with the Bayley-4 scores. In contrast, one study (Provost et al., 2004) shows that the relationship between the Bayley–II and the PDMS–2 were not as strong when comparing the Bayley–II PDI and the PDMS–2 in a sample of 12-month-old at-risk children and matched controls. In the at-risk group, the Bayley–II PDI was 65.6 and the PDMS–2 Total Motor Quotient was 83. For the matched controls, the Bayley–II PDI was 92.3, whereas the PDMS–2 Total Motor Quotient was 101.5.

Table 3. Comparison of Bayley-4, WPPSI–IV, and PDMS–2 Scores

Bayley-4	Score	WPPSI–IV	Score	Difference
Cognitive	98.5	FSIQ	103.3	-4.8
Cognitive	98.5	VSI	100.2	-1.7
Language	100.6	VCI	103.0	-2.4
Bayley-4	Score	PDMS–2	Score	Difference
Motor	99.5	Total Motor Quotient	99.5	0.0
Fine Motor	99.5	Fine Motor Quotient	97.5	2.0
Gross Motor	99.5	Gross Motor Quotient	101.1	-1.6

Clinical and Diagnostic Accuracy

The validity of the Bayley-4 scores can also be established using data comparing special groups to matched controls. For the Bayley-4, some special group studies were conducted including children with: Down syndrome (DS), autism spectrum disorder (ASD), language delay (LD), specific language impairment (SLI), developmental delay (DD), motor impairment (MI), those born moderate/late premature (MLP) and very/extremely premature (VEP), and prenatal alcohol and drug exposure (PDAE). The Cognitive, Language, and Motor Scales scores of these groups and the matched controls are found in Table 4 (Bayley & Aylward, 2019).

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

Table 4. Comparisons of Bayley-4 Special Groups and Matched Controls

Group	*N*	Cognitive	Language	Motor
Down syndrome	54	70.0	67.6	65.1
Matched controls		102.0	101.0	100.1
Difference		-32.0	-33.4	-35.0
ASD	31	72.0	62.8	71.4
Matched controls		97.0	97.9	94.9
Difference		-25.0	-35.1	-23.5
Language delay	25	92.0	79.9	90.0
Matched controls		104.5	107.0	103.0
Difference		-12.5	-27.1	-13.0
SLI	25	88.0	81.8	85.2
Matched controls		101.0	101.9	100.1
Difference		-13.0	-20.1	-14.9
Developmental delay	57	82.5	78.8	81.5
Matched controls		97.0	96.5	96.7
Difference		-14.5	-17.7	-15.2
Motor impairment	40	77.5	78.4	72.9
Matched controls		103.5	103.9	102.3
Difference		-26.0	-25.5	-29.4
MLP	70	92.5	93.2	89.7
Matched controls		103.0	101.8	102.0
Difference		-10.5	-8.6	-12.3
VEP	66	83.0	85.9	78.5
Matched controls		100.5	100.5	99.0
Difference		-17.5	-14.6	-20.5
PDAE	44	76.0	80.5	83.4
Matched controls		100.5	100.8	99.3
Difference		-24.5	-20.3	-15.9

In every case, the clinical group means were significantly below the matched Bayley-4 normative controls. In addition, the profiles of scores reflect a priori expectations. For example, with ASD, Language is the lowest score; in the DS group, global delays were found and the degrees of delay are compatible with the literature; with children born VEP, low scores are found in all three domains, particularly Motor, and these scores are lower than those obtained by the MLP group whose scores are low average, yet still significantly lower than those obtained by matched controls.

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

Conclusions

There is strong evidence to support the accuracy and validity of Bayley-4 scores. The fact that it yields similar scores to the Bayley–III, uses updated representative demographics of the normative sample, has scores that are comparable to those on the WPPSI–IV and PDMS–2, as well as the consistent special group data, supports the changes made in the Bayley–III and carried forward in the Bayley-4. The Bayley–II norms are likely too conservative because of age-based item sets, normative sample demographics that were less representative, and a questionable Flynn effect. The 13-year run of the Bayley–II may have unduly influenced our reference point regarding what is considered "normal," making it more conservative. The Bayley-4 may be "recalibrating" what is considered typical development.

To further validate these conclusions, given the previous findings of the Bayley–III in Australia, a full standardization of norms is in progress in that country, followed by a norm validation in the U.K. Validation and standardization studies in other European countries are also being considered. This is a high-stakes issue with far ranging implications (Hack, 2012) in terms of determining rates of neurodevelopmental impairment, correctly interpreting results of outcome studies, and deciding who receives services and who does not. Therefore, it is imperative that we reach a consensus regarding which version of the Bayley Scales scores is most appropriate.

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

References

Anderson, P. J., & Burnett, A. (2017). Assessing developmental delay in early childhood–concerns with the Bayley–III scales. *The Clinical Neuropsychologist, 31*(2), 371–381. doi:10.1080/13854046.2016.1216518

Anderson, P. J., De Luca, C. R., Hutchinson, E., Roberts, G., Doyle, L. W., & Victorian Infant Collaborative Group. (2010). Underestimation of developmental delay by the new Bayley–III scale. *Archives of Pediatrics & Adolescent Medicine, 164*(4), 352–356. doi:10.1001/archpediatrics.2010.20

Aylward, G. P. (2013). Continuing issues with the Bayley–III: Where to go from here. *Journal of Developmental & Behavioral Pediatrics, 34*(9), 697–701. doi:10.1097/DBP.0000000000000000

Aylward, G. P., & Aylward, B. S. (2011). The changing yardstick in measurement of cognitive abilities in infancy. *Journal of Developmental & Behavioral Pediatrics, 32*(6), 465–468. doi:10.1097/DBP.0b013e3182202eb3

Bayley, N. (1969). *Bayley Scales of Infant Development* [Measurement instrument]. San Antonio, TX: The Psychological Corporation.

Bayley, N. (1993). *Bayley Scales of Infant Development* (2nd ed.) [Measurement instrument]. San Antonio, TX: The Psychological Corporation.

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development* (3rd ed.) [Measurement instrument]. San Antonio, TX: Pearson.

Bayley, N., & Aylward, G. P. (2019). *Bayley Scales of Infant and Toddler Development* (4th ed.) *technical manual*. Bloomington, MN: NCS Pearson.

Bos, A. F. (2013). Bayley–II or Bayley–III: What do the scores tell us? *Developmental Medicine & Child Neurology, 55*(11), 978–979. doi:10.1111/dmcn.12234

Bratsberg, B., & Rogeburg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences, 115*(26), 6674–6678. doi:10.1073/pnas.1718793115

Connolly, B. H., Dalton, L., Smith, J. B., Lamberth, N. G., McCay, B., & Murphy, W. (2006). Concurrent validity of the Bayley Scales of Infant Development II (BSID–II) motor scale and the Peabody Developmental Motor Scale II (PDMS-2) in 12-month-old infants. *Pediatric Physical Therapy, 18*(3), 190–196.

Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn effect: A systematic literature review. *Intelligence, 59,* 163–169. doi:10.1016/j.intell.2016.10.002

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54*(1), 5–20. doi:10.1037/0003-066X.54.1.5

Folio, R. M., & Fewell, R. R. (2000). *Peabody Developmental Motor Scales* (2nd ed.) [Measurement instrument]. Austin, TX: Pro-Ed.

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

Gauthier, S. M., Bauer, C. R., Messinger, D. S., & Closius, J. M. (1999). The Bayley Scales of Infant Development-II: Where to start? *Journal of Developmental & Behavioral Pediatrics, 20*(2), 75–79. doi:10.1097/00004703-199904000-00001

Griffiths, R., & Huntley, M. (1996). *Griffiths Mental Development Scales-Revised: Birth to 2 years* [Measurement instrument]. Oxford, U.K.: Hogrefe.

Hack, M. (2012). Dilemmas in the measurement of developmental outcomes of preterm children. *The Journal of Pediatrics, 160*(4), 537–538. doi:10.1016/j.jpeds.2011.11.021

Hack, M., Taylor, H. G., Drotar, D., Schluchter, M., Cartar, L., Wilson-Costello, D., . . . Morrow, M. (2005). Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of extremely low birth weight children at school age. *Pediatrics, 116*(2), 333–341. doi:10.1542/peds.2005-0173

Johnson, S., Moore, T., & Marlow, N. (2014). Using the Bayley–III to assess neurodevelopmental delay: Which cut-off should be used? *Pediatric Research, 75*(5), 670–674. doi:10.1038/pr.2014.10

Kaufman, A. S., & Applegate, B. (1988). Short forms of the K-ABC Mental Processing and Achievement Scales at ages 4 to 12½ years for clinical and screening purposes. *Journal of Clinical Child Psychology, 17*(4), 359–369. doi:10.1207/s15374424jccp1704_10

Lennon, E. M., Gardner, J. M., Karmel, B. Z., Flory, M. J. (2008). Bayley Scales of Infant Development. In J. B. Benson & M. M. Haith (Eds.), *Language, memory, and cognition in infancy and early childhood* (pp. 37–48). San Diego, CA: Academic Press.

Lowe, J. R., Erickson, S. J., Schrader, R., & Duncan, A. F. (2012). Comparison of the Bayley–II Mental Developmental Index and the Bayley–III cognitive scale: Are we measuring the same thing? *Acta Paediatrica, 101*(2): e55–e58. doi:10.1111/j.1651-2227.2011.02517.x

Lynn, R. (2009). What has caused the Flynn effect? Secular increases in the development quotients of infants. *Intelligence, 37*(1), 16–24. doi:10.1016/j.intell.2008.07.008

McClain, C., Provost, B., & Crowe, T. K. (2000). Motor development of two-year-old typically developing Native American children on the Bayley Scales of Infant Development II motor scale. *Pediatric Physical Therapy, 12,* 108–113.

Milne, S., McDonald, J., & Comino, E. J. (2012). The use of the Bayley Scales of Infant and Toddler Development III with clinical populations: A preliminary exploration. *Physical & Occupational Therapy in Pediatrics, 32*(1), 24–33. doi:10.3109/01942638.2011.592572

Moore, T., Johnson, S., Haider, S., Hennessy, E., & Marlow, N. (2012). Relationship between test scores using the second and third editions of the Bayley Scales in extremely preterm children. *The Journal of Pediatrics, 160*(4), 553–558. doi:10.1016/j.jpeds.2011.09.047

NCS Pearson. (2008). *Bayley–III Technical Report* (Report No. 2). Bloomington, MN: NCS Pearson.

Copyright © 2019 NCS Pearson, Inc. All rights reserved.

Provost, B., Heimerl, S., McClain, C., Kim, N. H., Lopez, B. R., & Kodituwakku, P. (2004). Concurrent validity of the Bayley Scales of Infant Development II motor scale and the Peabody Developmental Motor Scales-2 in children with developmental delays. *Pediatric Physical Therapy, 16*(3), 149–156. doi:10.1097/01.PEP.0000136005.41585.FE

Robertson, C. M., Hendson, L., Biggs, W. S., & Acton, B. V. (2010). Application of the Flynn effect for the Bayley III scales. *Archives of Pediatrics & Adolescent Medicine, 164*(11), 1072–1073. doi:10.1001/archpediatrics.2010.199

Spencer-Smith, M. M., Spittle, A. J., Lee, K. J., Doyle, L. W., & Anderson, P. (2015). Bayley–III cognitive and language scales in preterm children. *Pediatrics, 135*(5), e1258–e1265. doi:10.1542/peds.2014-3039

Synnes, A. R., Anson, S., Arkesteijn, A., Butt, A., Grunau, R. E., . . . Whitfield, M. F. (2010). School entry age outcomes for infants with birth weight ≤ 800 grams. *The Journal of Pediatrics, 157*(6), 989–994. doi:10.1016/j.jpeds.2010.06.016

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin, 140*(5), 1332–1360. doi:10.1037/a0037173

Vohr, B. R., Stephens, B. E., Higgins, R. D., Bann, C. M., Hintz, S. R., Das, A., . . . Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. (2012). Are outcomes of extremely preterm infants improving? Impact of Bayley assessment on outcomes. *The Journal of Pediatrics, 161*(2), 222–228. doi:10.1016/j.jpeds.2012.01.057

Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence* (4th ed.) [Measurement instrument]. Bloomington, MN: NCS Pearson.

Weiss, L. G., Gregoire, J., & Zhu, J. (2016). Flaws in Flynn effect research with the Wechsler scales. *Journal of Psychoeducational Assessment, 34*(5), 411–420. doi:10.1177/0734282915621222

Copyright © 2019 NCS Pearson, Inc. All rights reserved.