Bayley–III Technical Report 2

Factors Contributing to Differences Between Bayley-III and BSID-II Scores

The *Bayley Scales of Infant and Toddler Development–Third Edition* (Bayley–III; Bayley, 2006) measures cognitive, language, motor, social-emotional, and adaptive development and is a revision of its predecessor, the *Bayley Scales of Infant Development—Second Edition* (BSID–II; Bayley, 1993). The Bayley–III is a technically sound instrument, with strong internal consistency, as well as test–retest stability.

In the course of developing the Bayley–III, a study was conducted in which the Bayley–III and the BSID–II were administered in counterbalanced order to 102 children ages 1 month to 42 months. A comparison of the Bayley–III and the BSID–II mean scores showed that the Bayley–III scores were approximately 7 points higher than the BSID–II Mental Development Index and Psychomotor Index scores. This finding is inconsistent with the Flynn effect (Flynn, 1984, 1987). This report examines factors that may contribute to the differences between the BSID–II and Bayley–III scores.

One possible explanation for the differences in scores, regardless of the Flynn effect, is the change in demographic characteristics of the samples. The BSID–II and Bayley–III normative samples are representative of the U.S. population in 1988 and 2000, respectively. The changes in demographic characteristics from 1988 to 2000, particularly parent education level (PEL), are noteworthy. In 1988, the proportion of children from parents with lower education levels (grade 12 or lower) was higher than in 2000. The proportion of children from various ethnic and racial backgrounds, as well as regions of the country, changed from 1988 to 2000. Gagnon and Nagle (2000) researched the effect of cultural differences and other secular changes (e.g., socioeconomic characteristics) on infant scores. Further research is required to determine how such changes in demographic characteristics may have contributed to children's performance from the second to third edition.

Improved Representativeness of the Normative Sample

Clinical cases constituted 5.8% of the Bayley–III normative sample. The BSID–II did not include any identified clinical cases in the normative sample. The clinical cases were included to make the Bayley–III sample more representative of the population and the full range of abilities within this age cohort (1–42 months). Including clinical cases in the sample is considered good psychometric practice that prevents truncated norms (McFadden, 1996). Other instruments that include clinical cases in the normative sample and are used to assess children in this age range are the *Preschool Language Scale–Fourth Edition* (PLS–

4; Zimmerman, Steiner, & Pond, 2002), *Wechsler Preschool and Primary Scale of Intelligence–Third Edition* (WPPSI–III; Wechsler, 2002), and *Peabody Picture Vocabulary Test–Third Edition* (PPVT–III; Dunn & Dunn, 1997).

Norming Methodology

The methodology of norms development has improved dramatically since 1992. For example, advanced methodology based on both the Classical Testing Theory (CTT) and Item Response Theory (IRT) were applied in the norms development of the Bayley–III. The growth curve and psychological theory were considered and the continuous norming method was used to establish more precise norms, which are reported in narrower age bands (Wilkins, Rolfhus, Weiss, & Zhu, 2005). As a result, the precision of Bayley–III as an assessment tool has significantly improved.

Evidence of Validity Based on Relationships to Other Variables

Comparisons to Other Tests

Validity studies comparing the Bayley–III with other tests do not show indications of Bayley–III standard score inflation. Bayley–III composite and subtest standard scores met theoretical expectations and are consistent with the results of the WPPSI–III, the PLS–4, and the *Peabody Developmental Motor Scales, Second Edition* (PDMS–2; Folio & Fewell, 2000). For example, the highest correlation between the Bayley–III and the WPPSI–III was between the Language composite and the VIQ ($r = .83$). There is very little difference between the Bayley–III Language composite mean and the WPPSI–III VIQ and FSIQ (98.9, 100.3, and 98.0, respectively). Details can be found in chapter 5 of the Bayley–III Technical Manual (Bayley, 2006).

Comparison of the Clinical Samples to the Normative Sample

The Bayley–III clinical studies show that the Bayley–III has high discrimination and good sensitivity and specificity in discriminating clinical cases from the normal cases. The Bayley–III clinical group scores are within the expected range that is consistent with the diagnosis. For example, the means for the sample with Down syndrome met theoretical expectations (e.g., they were two or more standard deviations below the means of the matched normative sample for all subtest and composite scores of the Bayley–III). Chapter 5 of the Bayley–III Technical Manual provides a detailed description.

Comparisons with Other Test Revisions

Recent revisions of instruments from other publishers are showing similar relationships between the current and previous editions. For example, Stockman (2000) notes that the PPVT–III produces higher test scores than the previous edition. Stockman posits that multiple factors are contributing to the discrepancy in scores between the two instruments, including changes in the nature of the task and changes in the characteristics of the normative sample (e.g., the inclusion of clinical groups; changes in

the proportion of low social class). A more recently published instrument, the *Battelle Developmental Inventory–Second Edition* (BDI–2; Newborg, 2005), shows a similar increase in test performance from the first edition to the second edition. Table 7.4 of the Examiner's Manual shows that the median percentile rank for the first edition (BDI; Newborg, Stock, Wnek, Guidubaldi, Svinicki, 1984), when compared with the second edition, produced standard score equivalents that were several points below those of the second edition (e.g., BDI percentile rank of 42 for the Total score converts to a standard score of 97; BDI–2 total standard score is 101.1).

Changes in the Flynn Effect

The Flynn effect, which is the phenomenon that population intelligence test scores increase over time, is typically ascribed to changes in population intelligence scores. More recent research on the Flynn effect revealed that the magnitude of the score change seems to be decreasing in general, and smaller changes in scores were observed among children than adults (Zhu & Tulsky, 1999). Because the Bayley scales were not intended as an intelligence measure per se, the Flynn effect may not apply here. A change in Bayley scores over time and test revision may not follow the trend that would be predicted by the Flynn effect.

Indications of Scores Lower Than Expected for the BSID–II

Research conducted after the BSID–II was published shows that children scored lower on the BSID–II than on the BSID (Bayley, 1969), though the BSID–II may have produced scores that were lower than expected. For example, a longitudinal study of drug-exposed children conducted by Schuler, Nair, and Harrington (2003) showed that young children's performance on the BSID–II was markedly lower than on the BSID.

The complexities in determining the appropriate item set on the BSID–II—particularly with a clinical population—can lead to administration of an item set that is below the age-appropriate start point. The basal and ceiling criteria sometimes can be met on several adjacent item sets, resulting in standard scores that range from several standard deviations below the mean to within the normal range (Alfonso, Russo, Fortugno, & Rader, 2005; Gauthier, Bauer, Messinger, & Closius, 1999; Washington, Scott, Johnson, Wendel, & Hay, 1998). Therefore, it is possible to administer an item set to a normal child that meets the basal and ceiling criteria and obtain a standard score that is well below normal.

Summary

The difference between the BSID–II and the Bayley–III scores exists. Bayley–III scores are consistent with other newly revised ability tests and show expected levels in various clinical groups. The Bayley–III provides scores that more precisely reflect performance of children from a contemporary sample that is representative of the population. The factors identified as likely contributors to the difference between the

BSID–II and Bayley–III scores require further research to identify their relative contributions and interactions, and related factors.

References

Alfonso, V. C., Russo, P. M., Fortugno, D. A., & Rader, D. E. (2005). Critical review of the Bayley Scales of Infant Development–Second Edition: Implications for assessing young children with developmental delays. *The School Psychologist, Spring*. 67–73.

Bayley, N. (1969). *Bayley scales of infant development*. San Antonio, TX: The Psychological Corporation.

Bayley, N. (1993). *Bayley scales of infant development—Second edition*. San Antonio, TX: The Psychological Corporation.

Bayley, N. (2006). *Bayley scales of infant and toddler development–Third edition*. San Antonio, TX: Harcourt Assessment, Inc.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test, third edition.* Circle Pines, MN: American Guidance Service.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171–191.

Folio, R. M., & Fewell, R. R. (2000). *Peabody developmental motor scales, second edition.* Austin, TX: Pro-Ed.

Gagnon, S. G., & Nagle, R. J. (2000). Comparison of the revised and original versions of the Bayley Scales of Infant Development. *School Psychology International, 21 (3)*, 293–305.

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 and Behavioral Pediatrics, 20 (2)*, 75–79.

McFadden, T. U. (1996). Creating language impairments in typically-achieving children: The pitfalls of "normal" normative sampling. *Language, Speech, and Hearing Services in the Schools, 27,* 3–9.

Newborg, J., Stock, J. R., Wnek, J., Guidubaldi, J., & Svinicki, J. S. (1984). *Battelle Developmental Inventory.* Itasca, IL: Riverside.

Newborg, J. (2005). *Battelle developmental inventory second edition examiner's manual.* Itasca, IL: Riverside.

Schuler, M. E., Nair, P., & Harrington, D. (2003). Developmental outcome of drug-exposed children through 30 months: A comparison of Bayley and Bayley–II. *Psychological Assessment, 15*(3), 435–438.

Stockman, I. J. (2000). The new Peabody picture vocabulary test–third edition: An illusion of unbiased assessment. *Language, Speech, and Hearing Services in Schools, 31*, 340–353.

Washington, K., Scott, D. T., Johnson, K. A., Wendel, S. & Hay, A. E. (1998). The Bayley Scales of Infant Development–II and children with developmental delays: A clinical perspective. *Journal of Developmental and Behavioral Pediatrics, 19 (5),* 346–349.

Wechsler, D. (2002). *Wechsler preschool and primary scale of intelligence for children—Third edition.* San Antonio, TX: The Psychological Corporation.

Wilkins, C., Rolfhus, E., Weiss, L., & Zhu, J. (April 2005). *A simulation study comparing inferential and traditional norming with small sample sizes.* Paper presented at the 2005 Annual Meeting of the American Educational Research Association, Montreal, Canada.

Zhu, J., & Tulsky, D. S. (1999). Can IQ gain by accurately quantified by a simple difference formula? *Perceptual and Motor Skills, 88*, 1255–1260.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool language scale–Fourth edition.* San Antonio, TX: The Psychological Corporation.