nature publishing group

Using the Bayley-III to assess neurodevelopmental delay: which cut-off should be used?

Samantha Johnson[1], Tamanna Moore[2] and Neil Marlow[2]

BACKGROUND: As the latest edition of the Bayley Scales (Bayley-III) produces higher scores than its predecessor (BSID-II), there is uncertainty about how to classify moderate–severe neurodevelopmental delay. We have investigated agreement between classifications of delay made using the BSID-II and Bayley-III.

METHODS: BSID-II Mental Development Index (MDI) and Bayley-III cognitive and language scales were administered in 185 extremely preterm (<27 wk) children. A combined Bayley-III score (CB-III) was computed. Agreement between delay classified using MDI scores <70 and various Bayley-III cut-offs was assessed.

RESULTS: Bayley-III cognitive and language scores were close to the normative mean and were higher than BSID-II MDI scores. Nineteen (10.2%) children had MDI <70. Bayley-III scores <70 significantly underestimated the proportion with MDI <70. Bayley-III cognitive *and* language scores <85 had 99% agreement with MDI <70 and underestimated delay by 1.1%. CB-III scores <80 had 98% agreement and produced the same proportion with delay.

CONCLUSION: Bayley-III cognitive *and* language scores <85 or CB-III scores <80 provide the best definition of moderate-severe neurodevelopmental delay for equivalence with MDI <70. CB-III scores have the advantage of producing a single continuous outcome measure but require further validation. The relative accuracy of both tests for predicting long-term outcomes requires investigation.

The Bayley Scales are the most frequently used tests in infant developmental assessment. The second edition of the test, the BSID-II (1), was widely used as an outcome measure in epidemiological studies and randomized controlled trials of infant interventions. The sound psychometric properties of the BSID-II Mental Development Index (MDI), a composite measure of nonverbal cognitive and language development, engendered much professional confidence and the MDI rapidly became the gold standard for assessing neurodevelopmental outcome (2).

In 2006, the third edition, the Bayley-III (3), was published which has separate cognitive and language scales. Although strong correlations are reported between MDI scores and Bayley-III cognitive and language scores (4,5), concerns have arisen over how to interpret test scores. Bayley-III scores are up to 10 points higher than MDI scores (3–6). Mean scores of control groups are of a similar magnitude above the normative mean (7,8) and those of clinical populations are higher than anticipated (4,5,8). Thus there is concern that the Bayley-III underestimates developmental delay using conventional cut-offs (4,6,8–11). However, as yet it is not clear whether the Bayley-III underestimates developmental delay or the BSID-II overestimated it.

These issues have significant implications for research and clinical practice. Children with developmental problems may not be identified using the Bayley-III and may thus fail to be referred for intervention. For research, the underestimation of delay leads to reduced statistical power in randomized controlled trials where sample sizes have been calculated using prevalence estimates obtained from studies using the MDI. Some groups have opted to raise Bayley-III SD-banded cut-offs by 15-points (1 SD) to retain power for primary outcomes and to identify children in need of intervention (12). It has also been suggested that raising the cut-off by 10-points maximizes agreement with the MDI (5). However, there still remains the problem of how to use Bayley-III scores to provide a relevant and practical composite outcome for identifying delay in research (13). To address these issues, we have investigated the predictive value of different Bayley-III cut-offs for classifying neurodevelopmental delay as measured by the BSID-II MDI in children born extremely preterm.

RESULTS

Descriptive statistics are shown in **Table 1**. The mean MDI score (93; SD 18) was 7-points lower than the normative mean and 19 (10.2%) children had neurodevelopmental delay. Mean Bayley-III scores were considerably higher and were close to the normative mean (cognitive 96, SD 14; language 103, SD 19; CB-III 100, SD 16). Bayley-III cognitive, language and CB-III scores were 3, 10, and 7 points higher than MDI scores, respectively. These results and correlations between test scores have been published previously (5). Using conventional SD-banded cut-offs, Bayley-III scores resulted in fewer children with

[1]Department of Health Sciences, University of Leicester, Leicester, UK; [2]Research Department of Academic Neonatology, Institute for Women's Health, University College London, London, UK. Correspondence: Samantha Johnson (sjj19@le.ac.uk)

Received 8 July 2013; accepted 16 October 2013; advance online publication 26 February 2014. doi:10.1038/pr.2014.10

Copyright © 2014 International Pediatric Research Foundation, Inc.

Table 1. Descriptive statistics for BSID-II and Bayley-III tests administered to 185 children born extremely preterm

Test	Mean (SD)	Range	Categorization of developmental delay, *n* (%)			
			Severe (score <55)	Moderate (score 55–69)	Mild (score 70–84)	None (score ≥85)
BSID-II						
MDI	93.1 (18.2)	49–124	13 (7.0%)	6 (3.2%)	28 (15.1%)	138 (74.6%)
Bayley-III						
Cognitive composite	96.1 (13.7)	30–125	2 (1.1%)	6 (3.2%)	12 (6.5%)	165 (89.2%)
Language composite	103.2 (19.3)	47–147	3 (1.6%)	10 (5.4%)	13 (7.0%)	159 (85.9%)
CB-III	99.7 (15.6)	39–129	3 (1.6%)	8 (4.3%)	14 (7.6%)	160 (86.5%)

MDI, Mental Development Index.
All scores are standardized scores with a normative mean of 100 (SD = 15). Delay is categorized using conventional SD-banded cut-offs (none ≥ −1 SD; mild −1 SD to −2 SD; moderate −2 SD to −3 SD; severe ≤3 SD).

moderate–severe delay and 10–14% more children in the average range (**Table 1**).

Table 2 shows the proportion of children with moderate-severe neurodevelopmental delay measured by the MDI and Bayley-III cut-offs. All Bayley-III scores <70 under-estimated the prevalence of delay by 2–7%; this difference was statistically significant for cognitive scores, cognitive and language scores and CB-III scores <70. Using Bayley-III scores <80, CB-III scores identified the same proportion with delay whilst cognitive scores alone and cognitive and language scores <80 significantly under-estimated the prevalence of delay. Taking scores <85, the proportion of children identified with delay using cognitive or cognitive and language scores differed by only 0.5–1.1%; however, language, CB-III and cognitive or language scores <85 significantly over-estimated the proportion with delay (**Table 2**).

Although the proportion of children with moderate–severe neurodevelopmental delay may be similar, this is not sufficient to determine the best cut point as agreement may be poor at the individual level. Predictive values were therefore calculated to assess diagnostic utility of the Bayley-III (**Table 3**). There were only marginal differences with the greatest agreement being for cognitive and language scores <85 (99% agreement) and CB-III scores <80 (98% agreement). In general, classifications with the poorest agreement were Bayley-III scores <70; cognitive and language scores <70 had poor sensitivity resulting in only 32% of children with MDI<70 being detected as delayed using the Bayley-III; sensitivity was 42% and 58% respectively for Bayley-III cognitive and language scores <70.

DISCUSSION

The results of this study show that, when using the Bayley-III to measure neurodevelopmental delay, classifications of cognitive and language scores <85 or CB-III scores <80 have the highest agreement with MDI scores <70 from the previous edition. These cut-offs resulted in a similar prevalence of children with neurodevelopmental delay and afforded the best prediction of an MDI <70 at the individual level.

There were only marginal differences in the predictive value of the various Bayley-III cut-offs assessed. All those in which cognitive and language scores were combined had >95% agreement. Such levels are to be expected where both tests measure the same underlying constructs. Selecting the most appropriate Bayley-III cut-off is therefore a pay-off between optimising the accuracy of prevalence estimates and the identification of delay at the individual level. Although cognitive and language scores <85 classified almost all children with delay who had MDI <70 (99%), it under-identified delay by 1.1%. In comparison, CB-III scores produced an identical prevalence estimate and were only 1.08% less accurate in their agreement. The differences are so subtle that these cut-offs may be used interchangeably when categorizing outcomes, and both will improve statistical power for primary outcomes compared with conventional SD-banded cut-offs. CB-III scores have an added advantage however, in providing a single continuous outcome measure.

Although there were only marginal differences between the cut-offs explored, Bayley-III scores <70 significantly under-estimated the prevalence of MDI <70. This was expected as, although there are moderate to strong correlations between MDI and Bayley-III scores (3–5), there is poor agreement between the two tests and scores are off-set by around 3–7 points. Of most concern is that this difference is not linear resulting in greater differences in scores at the lower end of the scale (5,7). This has major implications for clinical populations, to whom the test is most frequently applied, in terms of reduced prevalence estimates for neurodevelopmental delay and under-identification of children in need of intervention. Randomized studies of developmental care and parenting interventions have produced mixed results, but there appears to be benefits in terms of improved neurodevelopmental outcomes over infancy and the preschool years (14,15). If such improvements can be assured then the early identification of extremely preterm children in need of intervention is crucial, for which well-standardized developmental tests with appropriate cut-offs need to be applied. We believe that the results of this study will provide clinicians with information to aid in the interpretation of Bayley-III scores for these purposes.

Whether the Bayley-III underestimates developmental delay or the BSID-II MDI overestimated it remains an open question. However, converging evidence suggests the former

Copyright © 2014 International Pediatric Research Foundation, Inc.

Table 2. Proportion of children with moderate–severe neurodevelopmental delay classified using the BSID-II MDI and Bayley-III cognitive and language composite scores

	Score <70		Score <80		Score <85	
Test	*n* (%)	*P*[a]	*n* (%)	*P*[a]	*n* (%)	*P*[a]
BSID-II						
MDI	19 (10.3%)	—	—	—	—	—
Bayley-III						
Cognitive composite	8 (4.3%)	**0.001**	10 (5.4%)	**0.004**	20 (10.8%)	1.000
Language composite	13 (7.0%)	0.109	24 (13.0%)	0.063	26 (14.1%)	**0.016**
Cognitive or language composite	15 (8.1%)	0.289	24 (13.0%)	0.063	29 (15.7%)	**0.002**
Cognitive and language composite	6 (3.2%)	**<0.001**	10 (5.4%)	**0.004**	17 (9.2%)	0.500
CB-III	11 (5.9%)	**0.008**	19 (10.3%)	1.000	25 (13.5%)	**0.031**

MDI, Mental Development Index.
[a]*P* values for difference in proportions between MDI <70 and Bayley-III classifications of neurodevelopmental delay calculated using McNemar's test. Bold *P* values denote significance at $P<0.05$. All scores are standardized scores with a normative mean of 100 (SD = 15).

Table 3. Results for prediction of MDI scores <70 using different Bayley-III combinations and cut-offs

	Predictive values for BSID-II MDI <70 with 95% CI					
Bayley-III	Sensitivity	Specificity	PPV	NPV	Agreement[a]	Agreement rank[b]
Cut-off <85						
Cognitive <85	0.89 (0.65–0.98)	0.98 (0.94–1.00)	0.85 (0.61–0.96)	0.99 (0.95–1.00)	97.3%	3
Language <85	1.00 (0.79–1.00)	0.96 (0.91–0.98)	0.73 (0.52–0.88)	1.00 (0.97–1.00)	96.2%	7
Cognitive *or* language <85	1.00 (0.79–1.00)	0.94 (0.89–0.97)	0.66 (0.46–0.81)	1.00 (0.97–1.00)	94.6%	12
Cognitive *and* language <85	0.89 (0.65–0.98)	1.00 (0.97–1.00)	1.00 (0.77–1.00)	0.99 (0.95–1.00)	98.9%	1
CB-III <85	1.00 (0.79–1.00)	0.96 (0.92–0.99)	0.76 (0.54–0.90)	1.00 (0.97–1.00)	96.8%	6
Cut-off <80						
Cognitive <80	0.53 (0.29–0.75)	1.00 (0.97–1.00)	1.00 (0.66–1.00)	0.95 (0.90–0.97)	95.1%	11
Language <80	1.00 (0.79–1.00)	0.97 (0.93–0.99)	0.79 (0.57–0.92)	1.00 (0.97–1.00)	97.3%	3
Cognitive *or* language <80	1.00 (0.79–1.00)	0.97 (0.93–0.99)	0.79 (0.57–0.92)	1.00 (0.97–1.00)	97.3%	3
Cognitive *and* language <80	0.53 (0.29–0.75)	1.00 (0.97–1.00)	1.00 (0.66–1.00)	0.95 (0.90–0.97)	95.1%	10
CB-III <80	0.89 (0.65–0.98)	0.99 (0.95–1.00)	0.89 (0.65–0.98)	0.99 (0.95–1.00)	97.8%	2
Cut-off <70						
Cognitive <70	0.42 (0.21–0.66)	1.00 (0.97–1.00)	1.00 (0.60–1.00)	0.94 (0.89–0.97)	94.6%	12
Language <70	0.58 (0.34–0.79)	0.99 (0.95–1.00)	0.85 (0.54–0.97)	0.95 (0.91–0.98)	94.6%	12
Cognitive *or* language <70	0.68 (0.43–0.86)	0.99 (0.95–1.00)	0.87 (0.58–0.98)	0.96 (0.92–0.99)	95.7%	8
Cognitive *and* language <70	0.32 (0.14–0.57)	1.00 (0.97–1.00)	1.00 (0.52–1.00)	0.93 (0.88–0.96)	93.0%	15
CB-III <70	0.58 (0.34–0.79)	1.00 (0.98–1.00)	1.00 (0.68–1.00)	0.95 (0.91–0.98)	95.7%	8

CI, confidence interval; MDI, Mental Development Index; NPV, negative predictive value; PPC, positive predictive value.
[a]Overall agreement is calculated as the total proportion of true positive and true negative classifications. [b]Rank order of agreement with lower rank indicating higher agreement.

may be the case. The restandardization of a scale typically results in a decrease in mean scores on the new test which is attributed to the Flynn effect, a generalized increase in standardized test scores over time (16). Although this was found when the second edition of the Bayley Scales was normed (17), the expected drop in scores was not evidenced in the standardization of the Bayley-III. Indeed, we and others have shown that Bayley-III scores are higher than MDI scores (3–5,7). Furthermore, group mean scores for clinical populations known to be at high risk for developmental problems, such as the present population, are anchored around the normative mean, and mean scores for normative control groups are higher than the norm (5,8). It is plausible that the Bayley-III underestimates developmental delay since 10% of children in the standardization sample had established developmental problems. This proportion far exceeds that of the normal distribution and questions the validity of the standardization sample as a normative reference (3). Over-sampling of children with developmental problems ultimately has the effect of making the test "easier" and provides a mechanism whereby

Copyright © 2014 International Pediatric Research Foundation, Inc.

the Bayley-III may lead to under-estimation of developmental delay (18). Further support is provided by a recent study in which a Bayley-III cognitive score <80 at 22 mo had the best prediction of an IQ score <70 at 3 y of age in very preterm children (D.E. Creighton, S. Tang, J.E. Newman, A. Holub, R. Sauve, unpublished data). However, this was not compared with the MDI or the predictive validity of other Bayley-III scores assessed. To address this question definitively longitudinal studies are needed to determine the relative predictive validity of both editions of the test.

This is the largest study to date in which the BSID-II and Bayley-III has been administered in the same sample permitting a direct comparison of scores on both tests. The methodology used ensured that children's performance on all items on both tests was directly assessed, rather than estimated. As children who could not complete the two tests in succession were excluded the prevalence of moderate–severe neurodevelopmental delay in our extremely preterm children is actually higher than reported here. However, this does not negate the validity of our findings since the aim of this study was to compare the relationship between the two tests in the same group of children rather than assess neurodevelopmental outcomes. Exploring these issues in an extremely preterm cohort in which neurodevelopmental problems are common was fortuitous in highlighting the nonlinear off-set in scores (5). However, caution should be observed before applying these findings to other populations. Future studies should aim to replicate the current findings in other large clinical populations, such as infants born at more mature preterm gestations, and at different ages of assessment. Exploration of these issues in community-based samples and in larger samples of children with developmental delay would improve the power of future studies and are needed to determine the most appropriate cut-off for identifying long-term developmental problems.

Given the potential for improved clinical utility of the Bayley-III for identifying profiles of abilities across separable domains, future research is needed to address the psychometric properties of the test in response to examiners' concerns and to identify more appropriate clinical, and potentially population-specific, cut-off points. This should include an exploration of the psychometric properties of CB-III scores for identifying concurrent delays and predicting future outcomes. Crucially, the relative accuracy of BSID-II MDI and Bayley-III cut-offs for predicting long-term outcomes requires investigation in order to determine whether the MDI overestimates or the Bayley-III underestimates neurodevelopmental delay. The agreement between classifications of delayed motor development measured using both tests also requires investigation. The present findings should be only applied to Bayley-III motor scale scores with caution before such studies have been conducted.

Conclusions

To identify moderate–severe developmental delay and provide sufficient power for primary outcomes, Bayley-III cognitive and language scores <85 or CB-III scores <80 provide the best correspondence with MDI scores <70. Using CB-III scores has the added advantage of producing a single continuous outcome measure; future studies should therefore explore the psychometric properties of these scores. Follow-up of this cohort is needed to determine the accuracy of the Bayley-III for predicting long-term cognitive outcomes.

METHODS

Participants

All babies born 22–26 wk of gestation in England in 2006 and who survived to discharge were invited to participate in a follow-up assessment at 2–3 y of age corrected for prematurity (19). Of those alive at 2.5 y, 576 (55.3%) were formally assessed at 27–48 mo corrected age. In 225 (39.1%) children who were free of major neurosensory impairment and whose first language was English, an examiner administered the BSID-II MDI and the Bayley-III cognitive and language scales. Of these, 40 (18%) children were unable to cooperate for the time required to complete both tests. Thus, data from 185 children aged 29–41 mo corrected age (median 33 mo; median gestational age 25 wk, range 22–26 wk; 90 (40%) males) were included in the present study (**Figure 1**). Parents provided informed consent and the study was approved by the Northern and Yorkshire Research Ethics Committee.

Procedure and Measures

The cognitive and language scales of the Bayley-III were administered first and overlapping MDI items were scored simultaneously. Additional MDI item were administered after the Bayley-III. Examiners had >95% agreement with an experienced developmental psychologist across item scores on both Bayley-III scales indicating excellent inter-rater reliability. The protocol for this study has been published previously (5). The BSID-II MDI comprises items to assess cognitive and language development and yields a single age-standardized score (mean 100; SD 15; range 50–150). Children with scores below the basal limit were assigned a nominal score of 49 to reflect severely delayed development (20). The Bayley-III yields separate scores for cognitive (mean 100; SD 15; range 55–155) and Language (mean 100; SD 15; range 45–155) Scales. Extrapolated norms were used to quantify development for scores <55 on the cognitive scale (21). A CB-III score was calculated using the average of the child's cognitive and language scores. To classify moderate–severe neurodevelopmental delay on the criterion measure, the conventional cut-off of MDI scores <70

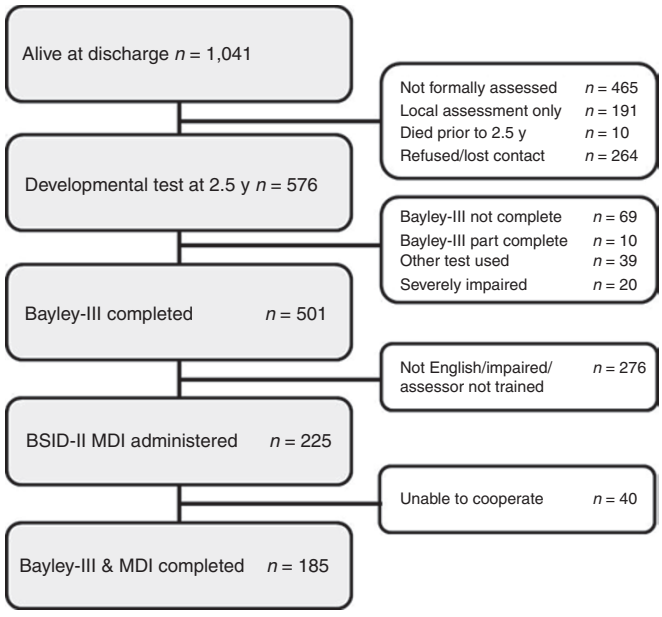

Figure 1. Participant recruitment and final study sample. MDI, Mental Development Index.

Copyright © 2014 International Pediatric Research Foundation, Inc.

was used. To compare agreement with different Bayley-III cut-offs, Bayley-III scores <80 and <85 were used for the 3 separate composite scores and for two combined measures: (i) cognitive or language score or (ii) a cognitive and language score below the cut-off.

Data Analyses

Data were double entered, verified and analyzed using IBM Statistics 20. Descriptive statistics for Bayley scores were calculated and differences in the proportion of children with developmental delay were analyzed using McNemar's tests. Classifications of moderate-severe delay using various Bayley-III cut points were cross-tabulated with MDI scores <70 and predictive values of sensitivity, specificity, positive predictive value and negative predictive value were calculated with 95% confidence intervals. The overall level of agreement between MDI scores <70 and each Bayley-III classification was computed by summing the proportion of true positive and true negative classifications. Agreement was ranked with lower ranks indicating better prediction.

STATEMENT OF FINANCIAL SUPPORT

This study was funded by the Medical Research Council, London, UK. Neil Marlow receives a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme at UCLH/UCL, London, UK.

Disclosure: The authors declare no conflict of interest.

REFERENCES

1. Bayley N. Bayley Scales of Infant Development. 2nd edn. San Antonio, TX: Psychological Corporation, 1993.
2. Johnson S, Marlow N. Developmental screen or developmental testing? Early Hum Dev 2006;82:173–83.
3. Bayley N. Bayley Scales of Infant and Toddler Development. 3rd edn. San Antonio, TX: Harcourt Assessment Inc, 2006.
4. Acton BV, Biggs WS, Creighton DE, et al. Overestimating neurodevelopment using the Bayley-III after early complex cardiac surgery. Pediatrics 2011;128:e794–800.
5. Moore T, Johnson S, Haider S, Hennessy E, Marlow N. Relationship between test scores using the second and third editions of the Bayley Scales in extremely preterm children. J Pediatr 2012;160:553–8.
6. Vohr BR, Stephens BE, Higgins RD, et al.; Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Are outcomes of extremely preterm infants improving? Impact of Bayley assessment on outcomes. J Pediatr 2012;161:222–8.e3.
7. Lowe JR, Erickson SJ, Schrader R, Duncan AF. Comparison of the Bayley II Mental Developmental Index and the Bayley III Cognitive Scale: are we measuring the same thing? Acta Paediatr 2012;101:e55–8.
8. Anderson PJ, De Luca CR, Hutchinson E, Roberts G, Doyle LW; Victorian Infant Collaborative Group. Underestimation of developmental delay by the new Bayley-III Scale. Arch Pediatr Adolesc Med 2010;164:352–6.
9. Msall ME. Overestimating neuroprotection in congenital heart disease: problems with Bayley III outcomes. Pediatrics 2011;128:e993–4.
10. Msall ME. Measuring outcomes after extreme prematurity with the Bayley-III Scales of infant and toddler development: a cautionary tale from Australia. Arch Pediatr Adolesc Med 2010;164:391–3.
11. Spittle AJ, Spencer-Smith MM, Eeles AL, et al. Does the Bayley-III Motor Scale at 2 years predict motor outcome at 4 years in very preterm children? Dev Med Child Neurol 2013;55:448–52.
12. Askie LM, Brocklehurst P, Darlow BA, Finer N, Schmidt B, Tarnow-Mordi W; NeOProM Collaborative Group. NeOProM: Neonatal Oxygenation Prospective Meta-analysis Collaboration study protocol. BMC Pediatr 2011;11:6.
13. Marlow N. Measuring neurodevelopmental outcome in neonatal trials: a continuing and increasing challenge. Arch Dis Child Fetal Neonatal Ed 2013;98:F554–8.
14. Moore T, Hennessy EM, Myles J, et al. Neurological and developmental outcome in extremely preterm children born in England in 1995 and 2006: the EPICure studies. BMJ 2012;345:e7961.
15. Johnson S, Wolke D, Marlow N. Outcome monitoring in preterm populations: Measures and methods. Zeitschrift fur Psychologie [J Psychol] 2008;216:135–46.
16. Pearson Assessment. Downward Extended Composite Normative Data for the Bayley Scales of Infant and Toddler Development-III. San Antonio, TX: NCS Pearson, 2005.
17. Spittle A, Orton J, Anderson P, Boyd R, Doyle LW. Early developmental intervention programmes post-hospital discharge to prevent motor and cognitive impairments in preterm infants. Cochrane Database Syst Rev 2012;12:CD005495.
18. Benzies KM, Magill-Evans JE, Hayden KA, Ballantyne M. Key components of early intervention programs for preterm infants and their parents: a systematic review and meta-analysis. BMC Pregnancy Childbirth 2013;13:Suppl 1:S10.
19. Flynn J. Searching for justice: the discovery of IQ gains over time. Am Psychol 1999;54:5–20.
20. Gagnon S, Nagle R. Comparison of the revised and original versions of the Bayley scales of infant development. Sch Psychol Int 2000;21:293–305.
21. Peña ED, Spaulding TJ, Plante E. The composition of normative groups and diagnostic decision making: shooting ourselves in the foot. Am J Speech Lang Pathol 2006;15:247–54.

Copyright © 2014 International Pediatric Research Foundation, Inc.