Validity of New Ballard Score until 7th day of postnatal life in moderately preterm neonates

K Sasidharan, S Dutta, A Narang

Department of Pediatrics, Postgraduate Institute of Medical Education and Research, Chandigarh, India

Correspondence to: Sourabh Dutta, Department of Pediatrics, PGIMER, Chandigarh 160012, India; sourabhdutta@yahoo.co.in

Accepted 14 February 2008
Published Online First 12 March 2008

ABSTRACT

Objective: The New Ballard Score (NBS) has been evaluated only until 96 h of age. We studied the validity and reliability of NBS for gestational age (GA) assessment on days 1, 5 and 7 of postnatal age (PNA).

Design and setting: This prospective, analytical study was conducted in a level III neonatal unit.

Patients: Neonates born at a GA of 29–35 weeks (based on accurate last menstrual period (LMP)) were eligible. Encephalopathy, malformations, and unstable vitals were exclusion criteria. LMP-based GA was the gold standard. NBS was assessed within 24 h of birth by one rater, and two raters assessed NBS on days 5 and 7. All were blinded to LMP and one anothers' ratings. Recruitment continued until >100 subjects were enrolled with ⩾25 in each LMP-based GA group: 29–30 weeks, 31–32 weeks, 33–34 weeks and 35 weeks.

Main outcome: Correlation of GA assessed on day 7 with gold standard.

Results: 129 neonates were studied. NBS-based GA on days 5 or 7 did not differ from the gold standard GA by more than 2 weeks in any subject. On day 7, NBS overestimated GA in 26.7% and underestimated GA in 19.8% cases; all discrepancies were ⩽2 wks. Compared to gold standard GA, the intra-class correlations (ICCs) of the gold standard GA and the NBS-based GA of the day 1 rater, day 5 rater and day 7 rater were 0.94, 0.94 and 0.92, respectively. ICCs for inter-rater reliability on day 5 and day 7 were 0.97 and 0.96, respectively. Compared to the day 1 rater's raw NBS, the ICCs of day 5 and day 7 raters' total scores were 0.98 and 0.97, respectively; of day 5 and day 7 raters' neurological scores were 0.98 and 0.97; and of day 5 and day 7 raters' physical scores were 0.92 and 0.88. All ICCs mentioned above had p values <0.001.

Conclusions: NBS is a valid and reliable clinical tool for GA assessment until day 7. It slightly overestimates the GA with increasing PNA. Neurological signs are more reliable than physical ones.

What is already known on this topic

- The New Ballard Score has been tested until 96 h of postnatal age and has been found valid with high inter-rater reliability.

What this study adds

- The New Ballard Score gives a valid and reliable assessment of gestational age until at least day 7 of life.
- The neurological component of the New Ballard Score is more reliable than the physical component of the score on day 7 of life.

Gestational age (GA) assessment is an important part of early neonatal examination. The most widely accepted scoring system for postnatal estimation of GA is the New Ballard Score (NBS).[1] The NBS can be used in extremely premature infants and in various ethnic groups.[2] If the difference between GA assessed by NBS and that calculated from the last menstrual period (LMP) is not greater than 2 weeks, it is the LMP GA which is assumed to be correct; and vice versa if the difference is greater than 2 weeks. Thus, a difference up to ±2 weeks can be ignored for clinical purposes.

The ideal postnatal age for NBS examination of infants at 26 weeks or more extends until 96 h of postnatal age.[1] However, the literature is silent on the validity and inter-rater reliability of NBS after 96 h of life.

There are situations where an accurate GA by LMP or early pregnancy ultrasound is not available, and the neonate's first contact with a paediatrician happens to occur only after 4 days of life. These situations are especially common in developing countries where two-thirds of deliveries, including preterm ones, are conducted by birth attendants at home.[3] Of the outborn, home-delivered babies admitted to our unit in the last 5 years, 43% had their first contact with a physician after 4 days of life (unpublished). Under such circumstances, the paediatrician is faced with a dilemma regarding the assessment of GA by NBS. Knowledge of the accurate GA influences the level of care provided, anticipation of medical problems, the interpretation of anthropometric measures and developmental capabilities.[4] It would greatly benefit paediatricians to know how useful the NBS is, when applied beyond 4 days of life. With this background, we studied the validity and inter-rater agreement of NBS on the 5th and the 7th days of postnatal age.

METHODS

This prospective study was conducted in a level III neonatal unit in a medical institute in Northern India. The unit conducts a neonatology super-specialty programme and has, at any time, eight neonatology fellows, who served as raters in this study. In the run-up to the main study, 1 month was spent on objectively testing inter-rater reliability. Neonates with gestational age of 29 to 35 weeks were examined by the neonatology

fellows within 96 h of life. It was decided a priori that the main study would commence only after an intra-class coefficient of correlation (ICC) for absolute agreement above 0.95 was separately obtained for the total NBS, the neurological score and the physical score.

In the main study, inborn premature babies who were antenatally assessed to have a GA of 29–35 weeks were eligible for enrolment. Antenatal GA assessment was based on an accurate record of the LMP, and this was assumed to be the gold standard GA. "Accurate record" meant that the mother was sure of the exact date of the LMP and pregnancy was confirmed soon after a missing period; she had been having regular menstrual periods prior to conception; and she had not conceived immediately after stopping hormonal contraceptives. Neonates, who had encephalopathy due to any cause in the first week of life, who had congenital malformations or who were unable to maintain their vital parameters (heart rate, oxygen saturation, non-invasive blood pressure, respiration and axillary temperature when nursed in the neutral thermal zone) within normal limits without support, were excluded.

The published NBS maturity-rating table has totalled scores expressed as 5-point increments and each of these incremental scores corresponds to an even-numbered GA. For scores that fell between the 5-point range we extrapolated the GA, using a method recommended by the author of the original article (J Ballard, personal communication). For example, if a score of 15 corresponded to 30 weeks and 20 to 32 weeks; for a score of 16 we assigned 30 weeks; and for scores of 17, 18 and 19 we assigned 31 weeks.

The initial assessment was carried out at approximately 24 h of life by the most senior neonatology fellow (henceforth called "day 1 rater"). This was the ideal method of postnatal GA assessment as per current NBS recommendations. On the 5th day (between 96 h and 120 h of life) the GA was assessed independently by two neonatology fellows (henceforth called "day 5 rater A" and "day 5 rater B", respectively). On the 7th day (between 144 h and 168 h of life) the GA was again assessed independently by two other neonatology fellows (henceforth called "day 7 rater A" and "day 7 rater B", respectively). The assessment of the raters "A" on each of the days was used for analysing the validity of NBS vis-à-vis the GA by LMP (gold standard) and the GA assessed by rater 1. The assessments of the second raters ("B") on days 5 and 7 were compared with raters A, respectively, for analysing the inter-rater agreement on the respective days. Never did the same rater examine the baby twice. They were selected from the available pool of neonatology fellows after excluding the following: the person who attended the delivery of the baby, the person who was posted in the area where the baby happened to be admitted, and the day 1 rater. All raters were blinded to the antenatally assessed GA and the assessed GA of all other examiners.

For NBS assessment, the baby was moved to a special warmed examination room. To conceal the identity of the baby from the rater, the mother was asked to not enter the examination room, the baby's identification badge was covered by an opaque band, and all other identification data were concealed. The LMP-based GA, birth weight, gender, and NBS assessments were recorded. For each assessment, the item-wise raw score, total physical score, total neurological score, total score and GA based on the total score were recorded.

Recruitment was planned to continue until at least 100 subjects had all five NBS evaluations each. This included at least 25 subjects in each of the following GA groups: 29–30 weeks, 31–32 weeks, 33–34 weeks, and 35 weeks. This was a sample size of convenience.

Neonates were enrolled after the parents had been explained about the nature of the study, had read an information sheet, and had provided written informed consent. The study did not entail any invasive or painful procedure or collection of sensitive information.

STATISTICAL ANALYSIS

Single-measure ICC with a one-way random effects model was used to calculate the level of agreement between raters for the physical scores, neurological score, total scores and assessed GA. Agreement on the item-wise scores was evaluated using linearly weighted kappa statistics. The kappa statistic was interpreted as: <0.20: poor; 0.21–0.40: fair; 0.41–0.60: moderate; 0.61–0.80: good; 0.81–1.00: very good.[5]

RESULTS

In the run-in period for testing inter-rater reliability, nine neonates were assessed. The ICC was 0.98 for both total score and physical score; and was 0.97 for neurological score (all p values <0.001).

Two hundred and thirty-six neonates were evaluated in the main study. Among them, 15 did not undergo the day 1 GA assessment (12 neonates had inaccurate GA based on LMP, two had encephalopathy on day 1, and one had ambiguous genitalia). Another 89 were excluded after the day 1 assessment was performed (74 neonates were unable to maintain vital parameters unsupported, and in 15 cases, all five NBS assessments could not be completed for logistic and administrative reasons).

Ultimately, 129 babies successfully completed the entire study. Of them, 25 subjects had a GA of 29–30 completed weeks by LMP, 38 had a GA of 31–32 weeks, 41 had a GA of 33–34 weeks, and 25 had a GA of 35 weeks. Male babies accounted for 55.8% of the study population. The mean (SD) birth weight was 1619.90 (316.11) g and ranged from 830 to 2482 g. The mean (SD) GA by LMP was 32.64 (1.90) weeks.

The frequency of subjects at each GA evaluated by the gold standard method and day 1 rater in comparison to raters on days 5 and 7 is enumerated in table 1. The day 1 rater overestimated the GA in 7% cases, underestimated it in 32.6% cases and accurately estimated it in 60.4% cases. Combining both the raters on day 5, the GA was overestimated in 11.6% cases, underestimated in 31.4% cases and accurately estimated in 57% cases. Combining both the raters on day 7, the GA was overestimated in 26.7% cases, underestimated in 19.8% cases and accurately estimated in 53.5% cases.

Since the difference between GA assessed by NBS and that assessed by LMP is considered to be clinically significant only if it exceeds 2 weeks, we calculated the number of subjects who had significant differences from the gold standard GA on days 1, 5 and 7 (table 2). There was no subject in whom the assessed GA by NBS differed from the LMP GA by more than 2 weeks on any of these days. The relationship of the difference in GA by LMP and day 7 rater A and the mean of the GA by LMP and day 7 rater A is also depicted as a Bland–Altman plot (fig 1). There was no case in which the difference in GA by the two methods exceeded 2 weeks.

The degree of agreement between the GA calculated from LMP and that assessed by the day 1 rater, and by raters A on days 5 and 7 is shown in table 3. The ICC of day 5 rater A was similar to the ICC of day 1 rater, while that of day 7 rater A was

Table 1 Frequency of subjects at each gestational age (GA)

		GA assessed by LMP (gold standard)							GA assessed by day 1 rater						
		29	30	31	32	33	34	35	29	30	31	32	33	34	35
Total number		**9**	**16**	**14**	**25**	**28**	**12**	**25**	**13**	**11**	**39**	**1**	**33**	**6**	**26**
GA assessed by day 1 rater	29	5	8												
	30	3	8												
	31	1		13	25										
	32			1											
	33					24	9								
	34					3	3								
	35					1		25							
GA assessed by day 5 rater A	29	6	9						9	6					
	30	3	7						4	5	1				
	31			13	25						38				
	32			1								1			
	33					20	6						24	2	
	34					7	6	1					9	4	1
	35					1		24							25
GA assessed by day 5 rater B	29	3	9						7	5					
	30	5	7						6	5	1				
	31	1		14	23					1	36	1			
	32				2						2				
	33					16	8						20	4	
	34					9	4						11	2	
	35					3		25					2		26
GA assessed by day 7 rater A	29	1	3						1	3					
	30	5	12						10	6	1				
	31	3	1	9	18				2	2	26	1			
	32			5	7						12				
	33					13	6						18	1	
	34					14	6						15	5	
	35					1		25							26
GA assessed by day 7 rater B	29	2	1							2	1				
	30	6	13	1					12	7	1				
	31	1	2	4	16				1	2	20				
	32			9	9						17	1			
	33					6	6						12		
	34					21	6						21	6	
	35					1		25							26

LMP, last menstrual period.

marginally lower. The degree of agreement between the GA assessed by day 1 rater and that by raters A on days 5 and 7 is also depicted (table 3). The ICC of the day 5 rater was higher than with the day 7 rater. The inter-rater reliability between two raters was calculated on days 5 and 7. The ICC was high on both days, with little difference between days. The level of agreement between the two raters on day 7 is displayed as a Bland–Altman plot (fig 2). The mean of the difference in GA assessed by the two raters is -0.9. In only six (4.5%) cases did the difference in GA exceed the bounds of ± 1.96 SD (0.93 to -1.12).

The raw total NBS, the physical score and neurological score given by the day 1 rater were tested for the level of agreement and degree of correlation with that given by day 5 rater A and day 7 rater A (table 4). The ICC was high both for the total NBS as well as for the neurological score (both in the range of 0.97 to 0.98), but lower for the physical score (0.91 on day 5 and 0.88 on day 7).

The total NBS, the physical score and neurological score given by the two raters on day 5 were tested for the level of agreement (table 4). The ICC was very high for the total score and the neurological score, but was 0.91 for the physical score. When the total NBS, the physical score and neurological score given by the two raters on day 7 were tested for the level of agreement

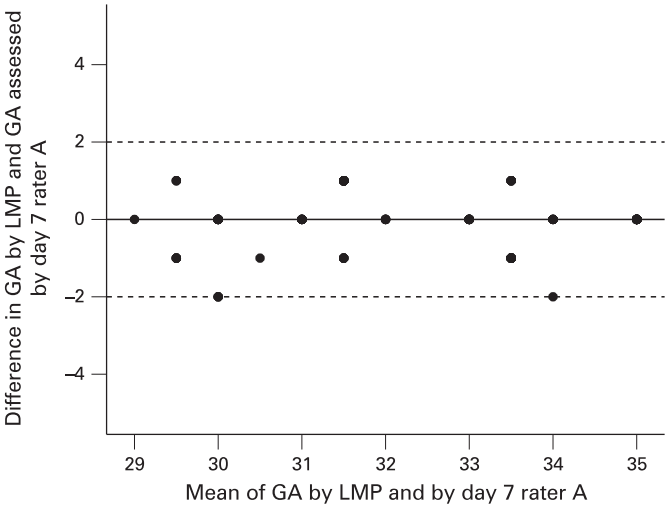

Figure 1 Bland–Altman plot of the difference in gestational age (GA) by last menstrual period (LMP) and day 7 rater A versus mean of GA by LMP and day 7 rater A. Solid horizontal line depicts ideal difference in GA by the two methods (ie, zero). Broken horizontal lines depict clinically permissible limits of difference in GA (± 2 weeks) Each dot may represent more than one subject.

Table 2 Differences between the gold standard GA and the GA assessed by NBS on days 1, 5 and 7

	Day 1 rater	Day 5 rater A	Day 7 rater A
No difference in GA	78	76	73
1 week difference in GA	49	52	52
2 weeks' difference in GA	2	1	4
>2 weeks' difference in GA	0	0	0

GA, gestational age; NBS, New Ballard Score.

and degree of correlation, the coefficients were similar to the corresponding figures on day 5.

The level of agreement of total NBS between the day 1 rater and day 7 rater A is depicted graphically as a Bland–Altman plot (fig 3). The mean of the difference in total NBS was −0.59. Only in five (4%) cases, did the difference exceed the bounds of ±1.96 SD (1.4 to −2.58). The level of agreement of total NBS between the two raters on day 7 is depicted graphically as a Bland–Altman plot (fig 4). The mean of the difference in total NBS was −0.19. Only in seven (5%) cases, did the difference exceed the bounds of ±1.96 SD (1.43 to −1.81).

Linear regression equations were generated to predict the day 1 NBS from the NBS on days 5 and 7. The equation for day 5 was day 1 NBS = 0.78 + 0.96 × NBS on day 5 (95% CI of the constant: 0.094 to 1.46 (p = 0.026) and 95% CI of the regression coefficient: 0.93 to 0.99 (p<0.001)). The R^2 of the model was 0.967 (p value <0.001). The equation for day 7 was day 1 NBS = −1 + 1.02 × NBS on day 7 (95% CI of the constant: −1.85 to −0.15 (p = 0.022) and 95% CI of the regression coefficient: 0.98 to 1.06 (p<0.001)). The R^2 of the model was 0.957 (p value <0.001).

Weighted kappa statistics were calculated for individual items (table 5). On day 5, the kappa statistic was "good" to "very good" for posture, heel to ear manoeuvre, lanugo, plantar surface and genitalia; and on day 7 the kappa statistic was "good" to "very good" for posture, plantar surface and genitalia. The kappa statistic for inter-rater reliability on day 5 was "good" to "very good" for posture, arm recoil, popliteal angle, scarf sign, heel to ear, lanugo, plantar surface and genitalia; and on day 7 was "good" to "very good" for posture, arm recoil, popliteal, heel to ear, skin, lanugo, plantar surface, breast, eye/ear and genitalia.

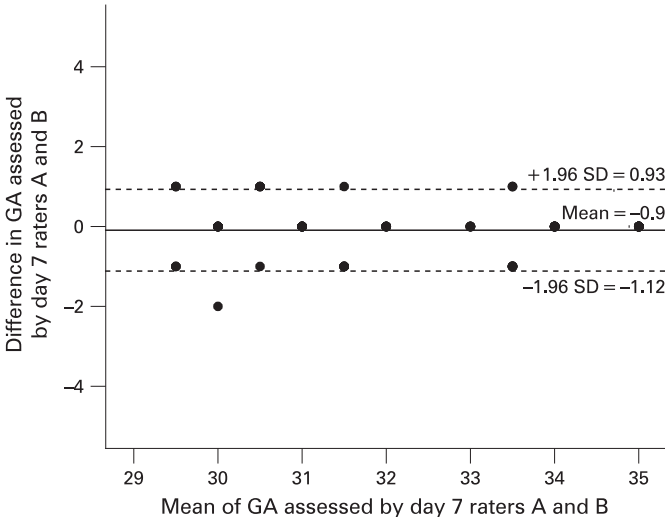

Figure 2 Bland–Altman plot of the difference in gestational age (GA) assessed by day 7 raters A and B versus mean of GA assessed by day 7 raters A and B. Solid horizontal line depicts the mean of difference in GA. Broken horizontal lines depict the 95% confidence limits of difference in GA. Each dot may represent more than one subject.

Table 3 Correlation of the assessed gestational age between various comparison groups

	Intra-class correlation	
Comparison groups	**Coefficient (95% CI)**	**p Value**
Gold standard vs day 1 rater	0.94 (0.92 to 0.96)	<0.001
Gold standard vs day 5 rater A	0.94 (0.92 to 0.96)	<0.001
Gold standard vs day 7 rater A	0.92 (0.89 to 0.95)	<0.001
Day 1 rater vs day 5 rater A	0.98 (0.97 to 0.98)	<0.001
Day 1 rater vs day 7 rater A	0.94 (0.92 to 0.96)	<0.001
Rater A on day 5 vs rater B on day 5	0.97 (0.96 to 0.98)	<0.001
Rater A on day 7 vs rater B on day 7	0.96 (0.94 to 0.97)	<0.001

DISCUSSION

This prospective observational study was conducted in babies of gestational age (by LMP) ranging from 29 weeks to 35 weeks to test the reliability of NBS in GA assessment on the 5th and 7th days of postnatal age. It is the first study of its kind and the results suggest that NBS can be used as a clinical tool for GA assessment at least until day 7 of postnatal age. Previous studies on gestation assessment have recommended lower postnatal ages, based on the inclusion criteria in their respective studies.[1 6 7]

The results of our study are applicable to neonates delivered from 29 to 35 weeks' GA. We chose this population for practical considerations. Neonates delivered at ⩽28 weeks in the community usually have problems for which they receive medical attention within 96 h of birth. On the other hand, neonates delivered at ⩾36 weeks' GA in our hospital generally remain well and it would be ethically unjustified to keep them in hospital solely for the study. The study population was recruited in four groups based on antenatally assessed GA to ensure representation from all gestational ages. The sample size was smaller than studies by Ballard *et al*, in which 252 and 578 babies, respectively, were enrolled.[1 6] However, this sample size was sufficient to address our study question. The mean (SD) GA by LMP (gold standard) in our study was 32.4 (1.8) weeks. If

Table 4 Correlation of the raw New Ballard Score values between various raters

	Intra-class correlation	
Comparison groups	**Coefficient (95% CI)**	**p Value**
Total score		
Day 1 rater vs day 5 rater A	0.98 (0.98 to 0.99)	<0.001
Day 1 rater vs day 7 rater A	0.97 (0.96 to 0.98)	<0.001
Rater A on day 5 vs rater B on day 5	0.98 (0.97 to 0.99)	<0.001
Rater A on day 7 vs rater B on day 7	0.98 (0.98 to 0.99)	<0.001
Neurological score		
Day 1 rater vs day 5 rater A	0.98 (0.97 to 0.99)	<0.001
Day 1 rater vs day 7 rater A	0.97 (0.96 to 0.98)	<0.001
Rater A on day 5 vs rater B on day 5	0.97 (0.96 to 0.98)	<0.001
Rater A on day 7 vs rater B on day 7	0.98 (0.97 to 0.99)	<0.001
Physical score		
Day 1 rater vs day 5 rater A	0.92 (0.88 to 0.94)	<0.001
Day 1 rater vs day 7 rater A	0.88 (0.83 to 0.91)	<0.001
Rater A on day 5 vs rater B on day 5	0.91 (0.88 to 0.94)	<0.001
Rater A on day 7 vs rater B on day 7	0.91 (0.88 to 0.94)	<0.001

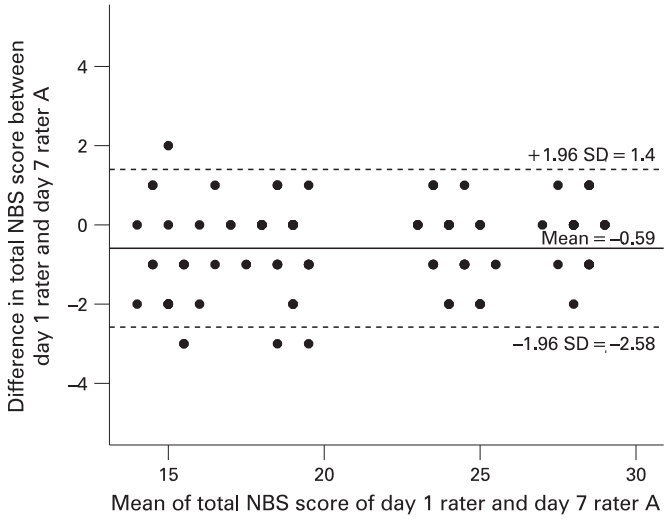

Figure 3 Bland–Altman plot of the difference in total New Ballard Score (NBS) of day 1 rater and day 7 rater A and mean of total NBS of day 1 rater and day 7 rater A. Solid horizontal line depicts mean of difference in total NBS. Broken horizontal lines depict 95% confidence limits of difference in GA. Each dot may represent more than one subject.

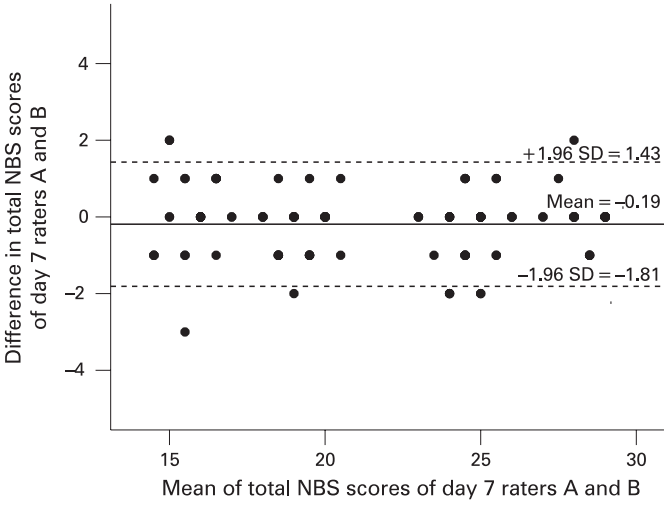

Figure 4 Bland–Altman plot of difference in total New Ballard Score (NBS) assessed by day 7 raters A and B versus mean of total NBS assessed by day 7 raters A and B. Solid horizontal line depicts mean of difference in total NBS. Broken horizontal lines depict 95% confidence limits of total NBS. Each dot may represent more than one subject.

one performs a post hoc power calculation, a sample size of 129 has greater than 99.9% power to detect a difference of 2 weeks or more from the GA by LMP with an alpha error of 5%.

Our study shows that NBS assessment on days 5 or 7 is as good as the assessment within the first 24 h of life. The GA calculated from LMP did not differ by more than 2 weeks from the GA assessed by NBS on day 1, on day 5 or on day 7, in any subject. Thus, NBS can be used as a clinical tool up to at least day 7 of postnatal life.

When exact agreement between the gold standard GA and GA assessed on days 1, 5 and 7 was analysed, some differences were found. The tendency to overestimate the GA increased from day 1 to day 5 to day 7. This phenomenon could be explained by the fact that the neonate would appear more mature with passage of time, and thus a day 5 assessment was likely to be more accurate than a day 7 assessment.

When the raw scores, broken down into neurological and physical scores were analysed, a high level of agreement for total NBS, neurological score and physical score was obtained between the day 1 rater and the raters on day 5 and day 7. However, on both days, the ICC for the physical scores was lower than that for the neurological scores. The inter-rater agreement between the two observers for physical score was lower than the neurological score on days 5 and 7. These findings suggest that physical features change more rapidly than neurological features postnatally and the rate of change in physical characteristics is not uniform.

In our study, posture was a parameter found to have excellent reliability on day 5 and day 7. This was consistent with the observation by Ballard *et al* that posture was the neurological characteristic with the highest coefficient of correlation (0.82) with antenatally assessed GA.[1] Among the physical characteristics, genitalia had the highest reliability. This was again consistent with previous observations that among physical characteristics, genitalia had the best coefficient of correlation (0.82).

Not only does the maturity of the skin increase rapidly after birth, it also varies widely with increasing postnatal age for the necessity of extrauterine adaptation.[8] This could be the reason for the lower inter-rater reliability of physical characteristics, which includes skin, lanugo, plantar surface, and breast, in GA assessment beyond 96 h of postnatal age. We hypothesise that the development of genitalia may not be affected to the same

Table 5 Weighted kappa statistics of item-wise scores

Items	Day 1 rater vs day 5 rater A (95% CI)	Day 1 rater vs day 7 rater A (95% CI)	Day 5 rater A vs day 5 rater B (95% CI)	Day 7 rater A vs day 7 rater B (95% CI)
Posture	0.86 (0.77 to 0.94)	0.83 (0.75 to 0.92)	0.91 (0.84 to 0.98)	0.89 (0.82 to 0.96)
Square window	0.54 (0.43 to 0.65)	0.48 (0.38 to 0.58)	0.58 (0.47 to 0.70)	0.45 (0.35 to 0.55)
Arm recoil	0.53 (0.43 to 0.63)	0.33 (0.2 to 0.45)	0.68 (0.59 to 0.77)	0.66 (0.56 to 0.76)
Popliteal angle	0.59 (0.47 to 0.71)	0.57 (0.45 to 0.69)	0.62 (0.50 to 0.74)	0.64 (0.52 to 0.76)
Scarf sign	0.6 (0.47 to 0.72)	0.41 (0.29 to 0.53)	0.79 (0.70 to 0.89)	0.57 (0.46 to 0.69)
Heel to ear	0.77 (0.69 to 0.86)	0.43 (0.32 to 0.54)	0.76 (0.67 to 0.85)	0.75 (0.64 to 0.85)
Skin	0.54 (0.37 to 0.7)	0.34 (0.16 to 0.52)	0.32 (0.14 to 0.49)	0.69 (0.55 to 0.83)
Lanugo	0.64 (0.51 to 0.76)	0.60 (0.48 to 0.73)	0.67 (0.55 to 0.79)	0.78 (0.68 to 0.88)
Plantar surface	0.64 (0.53 to 0.75)	0.61 (0.49 to 0.72)	0.70 (0.60 to 0.81)	0.72 (0.62 to 0.83)
Breast	0.51 (0.37 to 0.64)	0.39 (0.24 to 0.54)	0.32 (0.19 to 0.46)	0.67 (0.54 to 0.79)
Eye/ear	0.38 (0.24 to 0.53)	0.39 (0.25 to 0.52)	0.30 (0.16 to 0.44)	0.65 (0.54 to 0.77)
Genitalia	0.74 (0.62 to 0.86)	0.66 (0.52 to 0.80)	0.66 (0.53 to 0.80)	0.82 (0.70 to 0.94)

All p values <0.001.

extent by the extrauterine environment, which could explain the greater reliability of this parameter.

There were certain limitations of this study. Each GA group could not be analysed separately, because of the small numbers involved. We cannot exclude the possibility of an upper ceiling effect in our study, because of the way in which our subjects were recruited. An upper ceiling effect may inflate the magnitude of the ICC, especially at higher GA. Our raters were neonatology fellows, who, by virtue of their training programme, were well versed in NBS and whose high inter-rater reliability for the current standard of NBS assessment (ie, <96 h) was established before the study began. Since the study started only after high inter-rater reliability was confirmed, the results may not be generalisable into environments where similar well-trained observers are not available. Even something as basic as NBS assessment requires training and quality assurance. Thus, our study may have overestimated the true ICC values. Further studies are required to confirm whether our findings hold good with respect to the run-of-the-mill paediatrician.

This study demonstrates that the NBS can be reliably used until day 7 of postnatal age for clinical purposes, but the raw scores and individual item scores do show some discrepancies. This study has important implications, particularly in situations where newborns encounter physicians beyond the first days of life. Further studies are required to ascertain its applicability in sicker and more premature neonates and at higher postnatal days of life.

Competing interests: None.

Ethics approval: The study was approved by the Institute Ethics Committee.

REFERENCES

1. **Ballard JL,** Khoury JC, Wedig K, *et al*. New Ballard Score, expanded to include extremely premature infants. *J Pediatr* 1991;**119**:417–23.
2. **Feresu SA.** Does the modified Ballard method of assessing gestational age perform well in a Zimbabwean population? *Cent Afr J Med* 2003;**49**:97–103.
3. **Indian Institute of Population Studies**. National family health survey-II, 1998–99. 2000. Mumbai. http://www.nfhsindia.org/india2.html
4. **Kramer MS,** McLean FH, Boyd ME, *et al*. The validity of gestational age estimation by menstrual dating in term, preterm, and postterm gestations. *JAMA* 1988;**260**:3306–8.
5. **Landis JR,** Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;**33**:363–74.
6. **Ballard JL,** Novak KK, Driver M. A simplified score for assessment of fetal maturation of newly born infants. *J Pediatr* 1979;**95**:769–74.
7. **Farr V.** Estimation of gestational age by neurological assessment in first week of life. *Arch Dis Child* 1968;**43**:353–7.
8. **Rutter N.** The immature skin. *Eur J Pediatr* 1996;**155**(Suppl. 2):S18–20.

Validity of New Ballard Score until 7th day of postnatal life in moderately preterm neonates

K Sasidharan, S Dutta and A Narang

Arch Dis Child Fetal Neonatal Ed 2009 94: F39-F44
doi: 10.1136/adc.2007.122564