Early Human Development 170 (2022) 105610

Contents lists available at ScienceDirect

Early Human Development

journal homepage: www.elsevier.com/locate/earlhumdev

The stability of the Bayley scales in early childhood and its relationship with future intellectual abilities in a low to middle income country

Ingrid Kvestad [a,b,*], Mari Hysing [c], Suman Ranjitkar [d], Merina Shrestha [d], Manjeswori Ulak [d,e], Ram K. Chandyo [f], Tor A. Strand [b]

[a] Regional Centre for Child and Youth Mental Health and Child Welfare, NORCE Norwegian Research Centre, Bergen, Norway
[b] Department of Research, Innlandet Hospital Trust, Lillehammer, Norway
[c] Department of Psychosocial Science, Faculty of Psychology, University of Bergen, Bergen, Norway
[d] Department of Child Health, Institute of Medicine, Tribhuvan University, Kathmandu, Nepal
[e] Centre for Intervention Science in Maternal and Child Health, Centre for International Health, Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway
[f] Department of Community Medicine, Kathmandu Medical College, Kathmandu, Nepal

ARTICLE INFO

Keywords:
Bayley scales
Early childhood development
Stability
Predictive ability
Intellectual abilities

ABSTRACT

Background: The Bayley Scales of Infant and Toddler Development is widely used worldwide. The objective of the current study was to measure the stability of the Bayley Scales during early childhood and its relationship with intellectual abilities at four years in young Nepalese children.
Methods: In a prospective cohort we used the Bayley 3rd edition to measure early child development in 529 Nepalese children at 6–11, 18–23 and 30–35 months. At four years, we used the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) to measure intellectual abilities. We expressed the stability of the Bayley scores by intraclass correlation coefficients (ICCs) and concordance correlation coefficients (CCCs). The relationship between the Bayley scores and the WPPSI full-scale IQ (FSIQ) at four years was examined in regression models.
Results: The ICCs between the Bayley scores across timepoints were 0.01 (95 % CI -0.06, 0.04), 0.19 (95 % CI 0.15, 0.26) and 0.22 (95 % CI 0.17, 0.28) for the Cognitive, Language and Motor composite scores. The CCC for the composite scores ranged from 0.05 to 0.20 between 6 and 11 and 30–35 months and from 0.20 to 0.36 between 18 and 23 and 30–35 months. The Bayley scores at 6–11, 18–23 and 30–35 months explained 3 %, 20 % and 36 % of the variation of the FSIQ.
Conclusion: The stability of the Bayley scales is poor in early childhood, and its relationship with future intellectual abilities is poor in infancy but improves slightly with age in early childhood. Findings from this large community-based cohort of healthy at-risk children are relevant when measuring early child development worldwide.

1. Introduction

The Bayley Scales of Infant and Toddler Development (Bayley scales) is widely used as an assessment tool of early child development (ECD) worldwide. Although widely used and accepted, the Bayley scales have been criticized for its poor reliability and poor ability to predict future intellectual abilities [1,2].

Only a few studies have measured the stability of the Bayley scales throughout early childhood, showing poor and variable correlations between Bayley scores over time in full term [3,4] and preterm children [5]. Common for these studies was a low sample size (N < 100) and that all were done in high-income settings. An Indonesian study assessed the stability of the Bayley scales first version through repeated measurements in two cohorts showing only modest stability of the scales up to 18 months, with an increased stability after the second year of life [6]. More recent studies on the third version of the Bayley scales from low-to-middle income countries (LMICs) assess its usefulness and feasibility in a cross-cultural perspective [7–12]. None of these evaluate the stability of the Bayley scores through repeated measurements. The scarcity of studies from both high- and low-income countries, underline that the

* Corresponding author at: Regional Centre for Child and Youth Mental Health and Child Welfare, NORCE Norwegian Research Centre, Nygårdsgaten 112, 5008 Bergen, Norway.
E-mail address: inkv@norceresearch.no (I. Kvestad).

https://doi.org/10.1016/j.earlhumdev.2022.105610
Received 19 August 2021; Received in revised form 11 December 2021; Accepted 11 June 2022
Available online 16 June 2022
0378-3782/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

stability of the Bayley scales in early childhood needs further examination in studies with larger sample sizes.

A meta-analysis from 2013 on the predictive ability of the Bayley scales second edition in children born preterm demonstrated that the mental and motor development index explained 37 and 12 % of the variance in future cognitive function [13]. More recent studies in children born preterm not included in this meta-analysis, reported strong associations between the Bayley 3rd version at approximately 2 years and scores on the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) when the children were three and four years old [14,15]. Studies on the predictive ability of the Bayley scales from the general population, however, are scarce. A recent study in healthy Swedish children born to term, showed that there were only poor to moderate correlations between the Bayley scores measured at 2.5 years and WPPSI scores at 6.5 years [16]. From LMIC we have identified two studies addressing the predictive ability of the Bayley scales. In the previously mentioned Indonesian study, the Bayley scores showed no predictive power up to 18 months, increasing after 24 months of age [6], and a study from Bangladesh reported only modest correlations between the Bayley 2nd edition at 18 months and WPPSI scores at 61 months [17].

The first years of life from gestation and onwards is a period of rapid brain growth and development with fluctuating growth spurts [18]. Although genetically driven, brain development depends on biological and psychosocial influences. This rapid and fluctuating development as well as the increased susceptibility to external influences, leads to variability both within and between individuals in their developmental status. Consequently, the reliability and predictive ability of ECD measures may vary substantially according to age. Children from resource-poor populations in LMICs are subject to a range of risk factors. These risks may lead to increased fluctuations in their development, and thus also a risk for increased instability in the ECD measures. There are however a lack of studies addressing the stability and predictive ability of the Bayley scales specifically in a LMIC setting, and longitudinal studies examining this have been encouraged [19].

In Bhaktapur Nepal, we have followed a group of children originally included in a clinical trial, from 6 to 11 months old up to their 3rd birthday with Bayley 3rd version (Bayley-III) assessments at three time points and WPPSI-IV measured at approximately four years [20]. We have previously discussed the feasibility of the Bayley-III for the Nepalese setting demonstrating excellent measures of quality assurance [8]. These high-quality measures at multiple timepoints, provide a perfect opportunity to examine the stability and the predictive ability of the Bayley-III. Thus, the aim of the current study was to measure the stability of the Bayley-III subscale scores during early childhood, and to examine the relationship between these scores and the WPPSI-IV full-scale IQ measured at approximately four years in young Nepalese children.

2. Subjects and methods

2.1. Study setting and participants

We used data in 529 Nepalese children originally included to a community-based double-blind placebo-controlled trial measuring the effect of daily vitamin B12 supplementation for a year on neurodevelopment, growth, and anemia [20]. The trial was conducted from April 2015 until February 2018 and showed no effect of B12 supplementation on the main outcomes [21]. The study setting was Bhaktapur municipality close to the capital city Kathmandu. In the mother trial, we enrolled 600 mildly stunted (defined as a length-for-age < -1 z-score) children 6–11 months of age, from families that planned to reside in the area for the next 12 months, and with available informed consent from the caregivers. Children were excluded if they took supplements that contained vitamin B_{12}, had a severe systemic illness requiring hospitalization, if they were severely malnourished, severely anemic, or had ongoing infections that required medical treatment. After the completion of the clinical trial, children were followed with neurodevelopmental assessments every 12 months up to four years of age. From enrollment (6–11 months) to the end of the supplementation study (18–24 months), 26 children were lost to follow up (16 moved and 10 refused). From the end of supplementation to the 12 month follow up (30–35 months) 19 children were lost to follow up (12 moved and 7 refused), and from the 12 months follow up to 24 months follow up (42–47 months) 41 children were lost to follow up (27 moved and 14 refused) (Fig. 1). For the current analyses, we used data from the 529 children with neurodevelopmental measures at all assessment time points. The study received ethical clearance from the Nepal Health Research Council (NHRC, #233/2014, #73/2017) and from the Regional Committee for Medical and Health Research Ethics (REC # 2014/1528) in Norway. We obtained written informed consent from caregivers after providing thorough information on the study procedures.

2.2. Procedure

Field workers identified eligible children from immunization clinics or through home visits and children were enrolled at the study clinic by a physician or study supervisor. At enrollment, weight and length were taken, blood sample drawn, and caregivers were asked questions on child and family demographics. Socioeconomic status was assessed by a composite WAMI-index (range 0–1) to indicate wealth using variables

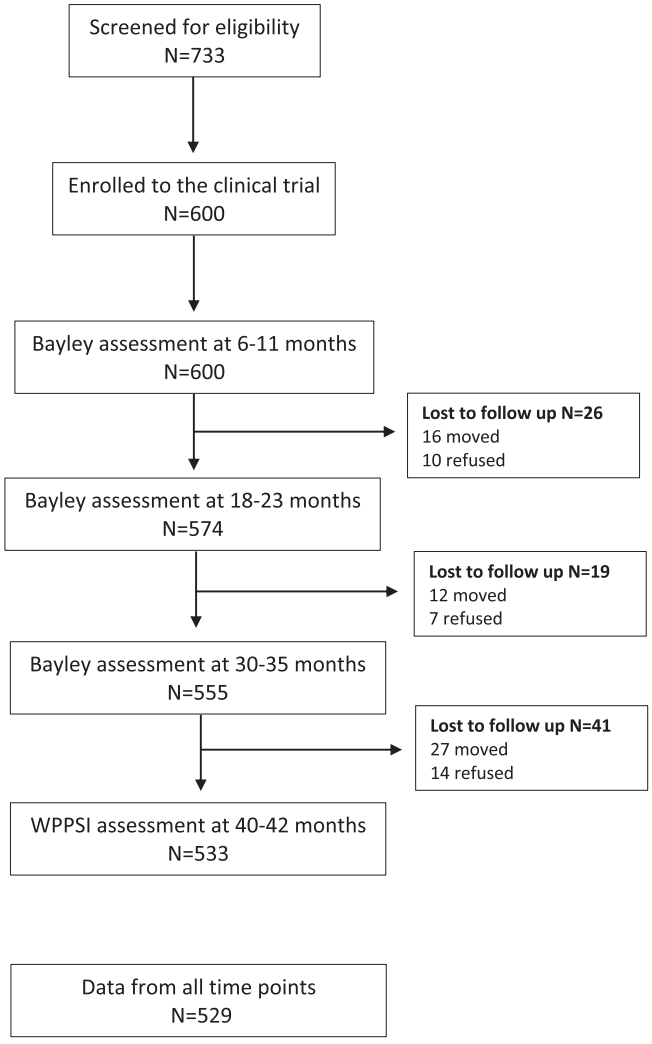

Fig. 1. Flow chart.

for water and sanitation access, household assets and maternal education [22].

2.3. Neurodevelopmental assessments

All neurodevelopmental assessments were done at the study clinic in well-lit rooms free from distractions by a team of three psychologists. At the three first time points we used the Bayley-III to assess neurodevelopment. The Bayley-III is a comprehensive assessment tool of neurodevelopment in infants and toddlers aged 1–42 months consisting of a Cognitive, Language (receptive and expressive), Motor (fine and gross), and Socio-emotional scale [23]. We converted the Bayley-III raw scores into subscale scores and composite scores based on the American norms [23]. For the current analyses, we used the Cognitive, Language and Motor composite scores (mean (SD) 100 (15)).

When the children were 42–47 months, we used the WPPSI-IV to assess intellectual abilities [24]. The WPPSI-IV is a widely used test to measure intellectual abilities in children between 2.6 and 7.3 years. In the present study, six subtests were included; Information, Receptive Vocabulary, Block Design, Picture Memory, Object Assembly and Zoo Locations to generate the Full-Scale IQ (FSIQ), the Verbal comprehension, Visuospatial, and the Working memory index. Raw scores were converted to index scores and IQ scores based on American norms [24]. For the current analyses we used the FSIQ (mean (SD) 100 (15)).

The translation and adaptation of the Bayley-III and its acceptability for a Nepalese context have thoroughly been described and discussed elsewhere [8]. Translation of the test instructions and some items in the language subtests were done according to standard procedures, and small adaptions to the materials (i.e., changing pictures and drawings such as vacuum cleaners, swimming pools and washing machines to more cultural appropriate items) were done to improve the acceptability for the Nepalese setting. Internal consistency of the scales ranged from poor to good with the poorest alpha values for the language subscales and better alpha values (>0.80) for the cognitive and gross motor subscales suggesting good internal consistency [8]. The distribution of the scores were similar to the American sample, except for the Language scales in which the scores were lower. For the WPPSI-IV we translated the general instructions for all subscales and also each item of the Receptive vocabulary and Information subscales according to standard procedures. We made no adaptions to the WPPSI-IV items and test materials for this study.

The assessor team were thoroughly trained in test procedures and performed standardization exercises in 20 children per outcome prior to the study assessments. In these standardization procedures, the assessors were required to reach an intra-class correlation coefficient (ICC) >0.90 with the expert rater (senior psychologist) [8]. During the study period, 7 % of the Bayley-III assessments and 10 % of the WPPSI-IV were double scored in a randomized manner by the expert rater, reaching an agreement of ICCs >0.95 for the Bayley-III assessments [8] and ICC > 0.98 for the WPPSI-IV.

2.4. Statistical analyses

Data are presented as means (SD) or numbers (%). The relationship between the Bayley composite scores at the three timepoints and between the Bayley composite scores and the WPPSI FSIQ was examined in a correlation matrix by Pearson correlations. We expressed the agreement between the Bayley-III composite scores at three timepoints (6–11, 18–23 and 30–35 months) by intraclass correlation coefficients (ICC), concordance correlation coefficients (CCC), and Cohen's kappa. The ICCs were calculated using one-way random effects models (Stata command "icc"). CCCs (Stata command "concord") were calculated between the composite scores at 6–11 and 30–35 months and between 18 and 23 and 30–35 months [25]. The Cohen's kappa (Stata command "kappa") was calculated for Bayley composite scores <70 across time points for each subscale. The agreement between the Bayley composite scores at these time points are also presented in Bland-Altman plots (stata command "loa") in supplementary files.

We used linear regressions to examine the relationship between the Bayley composite scores at the three time points and the FSIQ illustrated in fitted regression lines. From these models, we extracted the R-squared for each composite score at each time point. We also calculated the R-squared from a multiple linear regression model including all composite scores from the same time point. Scores <70 is a widely used cut off for cognitive delay both for Bayley composite scores and the FSIQ. Using kappa statistics (Stata command "kap"), we assessed the predictive ability of the Bayley composite scores <70 at each time point on FSIQ <70. In receiving operating characteristics (ROC) curve analyses (Stata command "rocgold"), we assessed the capability of the Bayley-III composite scores at the different time points to distinguish between FSIQ scores above and below cut off for cognitive delay. For these analyses, we defined cognitive delay both as FSIQ <70 and <−2 SDs below the mean score in the current sample (FSIQ scores <67.6). Although FSIQ<70 is the established cut off for cognitive delay, we also included −2 SD below the sample mean as a cut off since the WPPSI-IV has not been formally validated for a Nepalese context and there are no norms for this setting. The statistical analyses were done in Stata version 16 and JASP (0.14.0).

3. Results

Demographic characteristics of the 529 children are shown in Table 1. Children were on average 8.0 months at enrollment, with an even distribution of boys and girls. Approximately 20 % (104) were born at low birth weight, 11 % were born preterm, 33 % (174) were stunted (defined as length-for-age < −2 z- scores), and 20 % (104) were underweight. Half of the families owned land and 45 % of the families reported to live in rented houses. Table 2 shows the WPPSI FSIQ, and the three index mean (SD) scores. At 4 years, 25 (4.7 %) had FSIQ scores below 70, while 13 (2.5 %) of the children had FSIQ scores below 2 SDs (FSIQ scores <67.6) of the mean score of the current sample.

Agreement between Bayley-III subscale composite scores.

Table 3 shows the Bayley mean (SD) composite scores at the three timepoints and the ICCs (95%CI) for scores across timepoints and the

Table 1
Demographic characteristics of 529 Nepalese children.

	n (%)
Infant characteristics	
Age in month at enrollment, mean (SD)	8.0 (1.8)
Male child	271 (51.2)
Birthweight in grams, mean (SD)	2987.3 (1296.4)
Low birth weight (<2500 g)	104 (19.7)
Preterm birth (<37 weeks)	57 (10.8)
Demographic features	
Mothers' age, mean ± SD	27.7 (4.6)
Mothers who completed secondary school	342 (64.7)
Fathers who completed secondary school	344 (65.0)
Mothers who work	224 (42.3)
Fathers who work	515 (97.4)
Socio-economic status	
Wealth score (0–1)	0.62 (0.15)
Family stays in joint family	271 (51.2)
Family resides in rented house	238 (45.0)
Families with <3 rooms in their home	285 (53.9)
Family own land	264 (49.9)
Nutritional status of infants	
Underweight (weight for age z-score < −2)	104 (19.7)
Stunted (length for age z-score < −2)	174 (32.9)
Anemic (hemoglobin <11 g/dl)	339 (64.1)

SD – standard deviation.

Table 2
Mean (SD) and range of the Wechsler Preschool and Primary Scale of Intelligence, 4th edition (WPPSI-IV) full scale IQ and index scores in 529 Nepalese children.

WPPSI-IV subscales	Mean (SD)	Range	N (%) < 70	N (%) -2 SD
Full scale IQ	84.8 (8.6)	40–112	25 (4.7)	13 (2.5)
Verbal comprehension index	84.1 (7.9)	45–109		
Visuo-spatial index	85.6 (8.4)	45–118		
Working memory index	103.6 (13.3)	45–131		

SD – standard deviation, WPPSI-IV - Wechsler Preschool and Primary Scale of Intelligence, 4th edition.

Table 3
Mean (SD) and Intraclass correlation coefficients (ICCs)[1] Bayley Scales of Infant and Toddler Development, 3rd edition (Bayley-III) subscale composite scores and number (%) children with scores <70 at three time points in 529 Nepalese children.

	6–11 months	18–24 months	30–35 months	
	Mean (SD)	Mean (SD)	Mean (SD)	ICC[a] (95%CI)
Bayley-III subscale				
Cognitive composite	97.9 (10.5)	91.0 (7.7)	85.3 (6.9)	0.00 (−0.06, 0.04)
Language composite	85.6 (9.5)	93.2 (12.6)	95.3 (7.3)	0.19 (0.15, 0.26)
Motor composite	95.6 (12.8)	100.2 (8.6)	104.3 (9.9)	0.22 (0.17, 0.28)
	N (%)	N (%)	N (%)	Kappa
Cognitive composite <70	6 (1.1)	3 (0.6)	4 (0.8)	0.07
Language composite <70	24 (4.5)	26 (4.9)	5 (1)	0.08
Motor composite <70	12 (2.3)	2 (0.4)	2 (0.4)	0.15

Bayley-III - Bayley Scales of Infant and Toddler Development, 3rd edition; ICC – intraclass correlation; SD – standard deviation.
[a] One-way random effects models.

Cohen's kappas for composite scores <70. The Cognitive composite scores were the least stable with an ICC of 0.01 (95 % CI -0.06, 0.04), while the Language and Motor composite scores had ICCs of 0.19 (95 % CI 0.15, 0.26) and 0.22 (95 % CI 0.17, 0.28) respectively. The Cohen's kappas were 0.07 and 0.08 for the cognitive and language subscale respectively, and 0.15 for the motor subscale.

The relationship between the Bayley composite scores within and across time points are shown in Table 4 by Pearson correlation coefficients. The correlation coefficients between composite scores within the same age ranged from 0.28 (95%CI 0.20, 0.36) to 0.51 (95%CI 0.44, 0.57) and coefficients between domain composite scores across age varied from 0.11 (95%CI 0.02, 0.19) to 0.53 (95%CI 0.46, 0.59).

The CCCs between the domain composite scores at 6 and 11 and 30–35 months were 0.05 (95 % CI 0.01, 0.09), 0.13 (95 % CI 0.08,0.18) and 0.20 (95 % CI 0.14, 0.26) for the Cognitive, Language and Motor composites respectively, while the CCCs between 18 and 23 and 30–35 months were 0.20 (95 % CI 0.14, 0.26), 0.45 (95 % CI 0.40, 0.51), and 0.36 (95 % CI 0.29, 0.43) (Table 5). The difference in agreement between the domain composite scores are shown in Table 5 and in Bland-Altman plots (Supplementary Fig. 1).

The relationship between the Bayley-III subscale composite scores and WPPSI-IV FSIQ.

The correlation coefficients between the FSIQ and the Bayley composite scores ranged from 0.10 (95%CI 0.01, 0.18) to 0.52 (95%CI 0.45, 0.58) and are in general weak between the FSIQ and Bayley scores at 6–11 months reaching moderate strength between FSIQ and Bayley scores at 30–35 months (Table 4). Fig. 1 confirms the pattern from the correlation matrix, with fitted regression lines demonstrating the relationships increase with age. The regression models including all subscales explained 3 %, 20 % and 36 % of the variance in the FSIQ at 6–11, 18–23 and 30–35 months respectively (Fig. 2).

In ROC curves between the Bayley composite scores at the three timepoints and the cut off for delay when the children were four years, the area under the curve (AUC) were 0.64. 0.62 and 0.63 for the Cognitive, Language and Motor composite scores at 6–11 months using IQ < 70, and 0.56, 0.58 and 0.61 using <-2SD below sample mean. At 30–35 months, the AUCs were 0.86, 0.85 and 0.79 with <70 as cut off, and 0.93 and 0.82 and 0.86 using <-2SD below sample mean (Supplementary Fig. 2). The Cohen's kappas between the Bayley composite scores <70 at the different timepoints and the FSIQ <70 ranged from 0.05 to 0.26 (Table S1).

4. Discussion

In the current study, we measured the stability of Bayley scores during early childhood and its relationship with intellectual abilities at four years in 529 Nepalese children. The stability between the Bayley composite scores at the three time points were poor. The correlation coefficients between the same domains at different time points varied between 0.11 and 0.53, with lower coefficients at the earliest time points. It is important to keep in mind that these correlations correspond to explained variabilities of 1 to 25 %. The Bayley scales explained 3 %, 20 % and 36 % of the variability of the FSIQ at the three measurement time points. The agreement between having scores below the cut off for cognitive delay at each time point was poor both between the Bayley measures and between the Bayley and the FSIQ. The ability of the Bayley scales to discriminate between children with scores above and below cut offs for cognitive delay at four years was poor for the first Bayley measurements and improved with age.

Studies measuring the stability of the Bayley scales through repeated measurements in early childhood are scarce. Although an overall poor agreement taking all measurement time points into account, our findings suggest a pattern of poor correlations involving Bayley scores in infancy, and improved correlations with scores later in the period of early childhood. Overall, the variation in the coefficients in the current study is similar to previous studies [3–5]. A Polish study investigating the stability of scores on the Bayley-II taken at 12, 24 and 36 months, showed low to moderate correlations with stronger correlations between the latter time points than the first in accordance with current findings [26]. Using the Bayley-I, a previous study in Indonesia showed that the modest stability up to 18 months improved when the children reached 24 months [6]. When discussing the stability of ECD measures, it should be taken into account that since characteristics of the developing brain leads to variability both within and between individuals in their developmental sequences, a perfect agreement is not likely [18]. The fluctuations in the development decrease with time however [18], and the increased stability of scores with age seen in the current and previous studies is therefore expected. Hence, although poor agreement between scores taking all measurement time points into account, measurements taken later in early childhood should be considered more reliable.

Although the strengths of the relationship between the Bayley scores and the FSIQ are weak to moderate overall, results show a gradual improvement in the relationship between Bayley scores in the period of early childhood and intellectual abilities at four years. While the explained variance of 2 % using the Bayley scores at 6–11 months is considerably below what is described in the previous meta-analysis in preterm children, the explained variance of 36 % using measures at 30–35 months are in accordance with the findings of 37 % explained variability [13]. The gradual improved relationship between the Bayley scores and intellectual abilities, are complemented by the ROC analyses and kappa statistics. In the ROC analyses, the poor diagnostic prediction of the Bayley scores at the first time points, improved to good at the third

Table 4
Correlation coefficients (95 % confidence interval)[a] between Wechsler Preschool and Primary Scale of Intelligence, 4th edition full scale IQ score (FSIQ) at 4 years and Bayley Scales of Infant and Toddler Development, 3rd edition composite scores at three time points during early childhood in 529 Nepalese children.

		42–47 months	6–11 months			18–23 months			30–35 months		
		FSIQ	Cognitive	Language	Motor	Cognitive	Language	Motor	Cognitive	Language	Motor
42–47 months	FSIQ[b]										
6–11 months	Cognitive	0.10*	–								
		0.01–0.18	–								
	Language	0.16***	0.28***	–							
		0.07–0.24	0.20–0.36	–							
	Motor	0.14**	0.50***	0.37***	–						
		0.22–0.05	0.43–0.56	0.29–0.44	–						
18–23 months	Cognitive	0.32***	0.19***	0.18***	0.21***	–					
		0.24–0.39	0.11–0.27	0.10–0.26	0.13–0.29	–					
	Language	0.39***	0.19***	0.26***	0.19***	0.35***	–				
		0.31–0.46	0.10–0.27	0.18–0.34	0.11–0.27	0.27–0.42	–				
	Motor	0.29***	0.25***	0.24***	0.32***	0.35***	0.39***	–			
		0.21–0.36	0.17–0.33	0.15–0.31	0.24–0.39	0.27–0.42	0.31–0.46	–			
30–35 months	Cognitive	0.49***	0.11*	0.08	0.14**	0.26***	0.31***	0.29***	–		
		0.43–0.56	0.02–0.19	0.00–0.17	0.05–0.22	0.18–0.34	0.24–0.39	0.21–0.37	–		
	Language	0.52***	0.19***	0.22***	0.21***	0.28***	0.53***	0.36***	0.47***	–	
		0.45–0.58	0.10–0.27	0.13–0.30	0.12–0.29	0.20–0.36	0.46–0.59	0.28–0.43	0.40–0.54	–	
	Motor	0.39***	0.19***	0.18***	0.26***	0.27***	0.31***	0.40***	0.41***	0.51***	–
		0.32–0.46	0.11–0.27	0.09–0.26	0.18–0.34	0.19–0.35	0.23–0.35	0.32–0.47	0.33–0.48	0.44–0.57	–

[a] Pearson product moment correlation coefficient.
[b] Full scale intelligence quotients.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

Table 5
Concordance correlation coefficients (CCC), difference in average and 95 % limits of agreement between Bayley Scales of Infant and Toddler Development, 3rd edition (Bayley-III) composite scores at 6−11 and 30–35 months and 18–23 and 30–35 months in 529 Nepalese children.

	6–11 and 30–35 months			18–23 and 30–35 months		
Bayley-III composite scores	CCC (95%CI)	Difference Average	95 % Limits of Agreement	CCC (95%CI)	Difference Average	95 % Limits of Agreement
Cognitive	0.05 (0.01, 0.09)	12.62	−10.81, 36.05	0.20 (0.14, 0.26)	5.75	−11.70, 23.19
Language	0.13 (0.08, 0.18)	−9.68	−30.59, 11.23	0.45 (0.40, 0.51)	−2.05	−23.09, 18.99
Motor	0.20 (0.14, 0.26)	−8.69	−36.10, 18.73	0.36 (0.29, 0.43)	−4.09	−24.10, 15.93

Bayley-III - Bayley Scales of Infant and Toddler Development, 3rd edition; CCC – concord correlation coefficient.

year. Our findings are in accordance with a previous study that found very good prediction at 30 months of age, and lower prediction 8 and 18 months [27]. In another study in a LMIC setting, scores before approximately 24 months had low predictive power, while prediction increased after that age [6,17]. The Cohen's kappas demonstrate the poor predictive value of having poor Bayley scores in early childhood for poor FSIQ at four years. Due to the rapid and qualitative changes in the developing brain, a perfect prediction between the Bayley scales and later intellectual abilities should not be expected [1]. Notably, in the current study, measures taken at approximately 3 years showed better predictive ability than measures taken early.

The weak relationship between the Bayley-III and later intellectual abilities could partly be related to the characteristics of the study population. A previous study found lower correlations for infants born full-term than pre-term [15], and although there is a high level of premature born infants in the present sample, the majority is born to term. The present study including mildly stunted infants constitute a high-risk sample in the current population, however. It is thus reasonable to assume that these infants are subject to a range of risks that could affect their developmental trajectories over time [28,29]. As a consequence, the stability of the Bayley scores and its predictive ability might be lower in this sample than for children in more protective environments [5].

Our findings suggesting that Bayley scores in the period of early childhood fluctuate and have poor predictive ability have several clinical implications. For instance, Bayley measures from a single assessment time point should not be used alone to diagnose cognitive delay. Multiple Bayley measures to track developmental trajectories and information on the child's development from several sources is preferable to set diagnoses with certainty. Moreover, our findings indicate that the predictive ability of the Bayley test improve with age within the period of early childhood. Thus, higher confidence should be given to measures taken in older children than in younger, and if possible, WPPSI should be the preferred measure.

Our findings are also highly relevant for research in resource-poor settings in LMIC. The last decades there has been an increased awareness towards the risk of more extreme biological and psychosocial influences for young children in resource-poor circumstances leading to a loss of developmental potential [30]. Efforts to identify modifiable risk factors and to evaluate effective intervention strategies are accordingly called for. Our findings suggesting stability and predictive ability of the Bayley scores are poor in early life increasing in the period of early childhood, have important implications for the design of such studies. When reliability of an outcome measure decreases the unexplained variability increases, which has implications for the precision of effect measures estimates (p-values and confidence interval increase, and statistical power decrease). Such increased variability does not, however, affect the effect measure estimates. The unexplained variability of ECD measures in early childhood should be considered when estimating the samples size of a study. Moreover, although the relationship between the Bayley scores and the measure of intellectual abilities at 4 years improve within the period of early childhood, the strength of the estimates remains weak to moderate, and therefore care should be taken if study objectives involve prediction of future intellectual abilities. Measures taken later in the period of early childhood are preferable showing

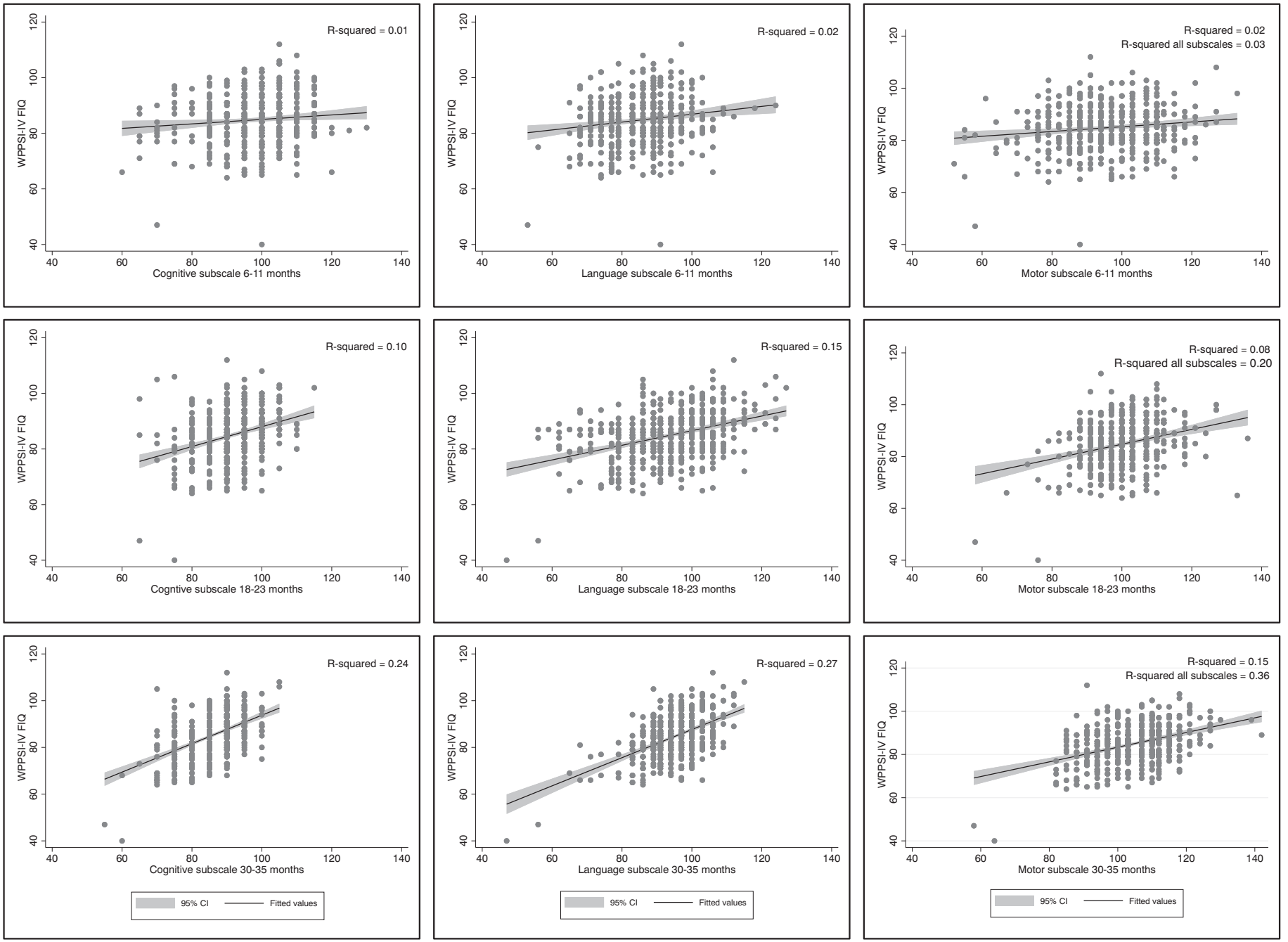

Fig. 2. Fitted regression lines between Bayley Scales of Infant and Toddler Development, 3rd edition composite scores at three timepoints during early childhood and Wechsler Preschool and Primary Scale of Intelligence, 4th edition Full scale IQ at 4 years in 529 Nepalese children.

that a longitudinal design with longer follow up time may be required.

Strengths of our study is the high-quality measurements with excellent inter-rater agreement [8] reducing random measurement errors, and the large sample size with repeated measurements and low attrition rate. Limitations include the use of neurodevelopmental assessment tools not formally validated for the Nepalese context and the use of US norms, as well as the high-risk sample of mildly stunted infants. The lack of validation does not, however, influence the internal validity of our findings but compromises its generalizability. In the interpretation of the current results, the lack of formally validated tests, the Nepalese context and the inclusion and exclusion criteria for the original trial need to be taken into consideration.

To conclude, our data suggest low stability of the Bayley scales early in infancy improving within the period of early childhood. The associations between the Bayley scales and later intellectual abilities also strengthen with age. Our large sample size and repeated high-quality measures of ECD in a community-based cohort provide strengths to our findings. Findings are of relevance for clinical practice and for the planning of research designs.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.earlhumdev.2022.105610.

Data sharing statement

Data available on request. In order to meet ethical requirements for the use of confidential patient data, requests must be approved by the Nepal Health Research Council (NHRC) and the Regional Committee for Medical and Health Research Ethics in Norway. Requests for data should be sent to the authors, by contacting NHRC (http://nhrc.gov.np), or by contacting the Department of Global Health and Primary Care at the University of Bergen (post@igs.uib.no).

Funding source

This work was supported by Thrasher Research Fund (award # 11512), the GC Rieber foundation, the University of Bergen (UoB), and the Research Council of Norway through a grant to Centre for Intervention Science in Maternal and Child Health (CISMAC).

CRediT authorship contribution statement

Ingrid Kvestad: funding acquisition, supervision, methodology, conceptualization, formal analysis, writing – Original draft. **Tor A. Strand:** project administration, funding acquisition, supervision, methodology, conceptualization, formal analysis, writing – Original draft. **Mari Hysing:** funding acquisition, supervision, methodology, conceptualization, writing – Original draft. **Suman Ranjitkar:** methodology, investigation, supervision, writing – review and editing. **Merina Shrestha:** methodology, supervision, writing – review and editing. **Manjeswori Ulak:** methodology, supervision, writing – review and editing. **Ram K. Chandyo:** project administration, funding acquisition, supervision, methodology, writing – Review and editing.

Declaration of competing interest

None.

Acknowledgement

We thank the psychologists Jaya Silpakar and Roshan Sintakala who performed the cognitive tests, supervisors, data management team and all field workers. We are grateful to all participant children and their families for their valuable time to participate in the study. We are also grateful to Shyam Sunder Dhaubhadel and all staff at Siddhi Memorial Hospital.

References

[1] P.J. Anderson, A. Burnett, Assessing developmental delay in early childhood—concerns with the Bayley-III scales, Clin. Neuropsychol. 31 (2) (2017) 371–381.
[2] G.P. Aylward, Continuing issues with the Bayley-III: where to go from here, Journal of Developmental & Behavioral Pediatrics. 34 (9) (2013) 697–701.
[3] M.T. Krogh, M.S. Væver, A longitudinal study of the predictive validity of the Bayley-III scales and subtests, Eur. J. Dev. Psychol. 16 (6) (2019) 727–738.
[4] M.A. Lobo, D.A. Paul, A. Mackley, J. Maher, J.C. Galloway, Instability of delay classification and determination of early intervention eligibility in the first two years of life, Res. Dev. Disabil. 35 (1) (2014) 117–126.
[5] M.M. Greene, K. Patra, J.M. Silvestri, M.N. Nelson, Re-evaluating preterm infants with the Bayley-III: patterns and predictors of change, Res. Dev. Disabil. 34 (7) (2013) 2107–2117.
[6] E. Pollitt, N. Triana, Stability, predictive validity, and sensitivity of mental and motor development scales and pre-school cognitive tests among low-income children in developing countries, Food Nutr. Bull. 20 (1) (1999) 45–52.
[7] J. Hua, Y. Li, K. Ye, Y. Ma, S. Lin, G. Gu, et al., The reliability and validity of Bayley-III cognitive scale in China's male and female children, Early Hum. Dev. 129 (2019) 71–78.
[8] S. Ranjitkar, I. Kvestad, T.A. Strand, M. Ulak, M. Shrestha, R.K. Chandyo, et al., Acceptability and reliability of the Bayley scales of infant and toddler development-III among children in BhaktapurNepal, Frontiers in Psychology 9 (2018).
[9] S. Manandhar, S. Dulal, D. Manandhar, N. Saville, A. Prost, Acceptability and reliability of the Bayley scales of infant development III cognitive and motor scales among children in Makwanpur, J. Nepal Health Res. Counc. 32 (2016) 47–50.
[10] C. Hanlon, G. Medhin, B. Worku, M. Tomlinson, A. Alem, M. Dewey, et al., Adapting the Bayley scales of infant and toddler development in Ethiopia: evaluation of reliability and validity, Child Care Health Dev. 42 (5) (2016) 699–708.
[11] N. Azari, F. Soleimani, R. Vameghi, F. Sajedi, S. Shahshahani, H. Karimi, et al., A psychometric study of the Bayley scales of infant and toddler development in Persian language children, Iran. J. Child Neurol. 11 (1) (2017) 50.
[12] M.S. McHenry, E. Oyungu, Z. Yang, A.C. Hines, A.R. Ombitsa, R.C. Vreeman, et al., Cultural adaptation of the Bayley scales of infant and toddler development, for use in Kenyan children aged 18–36 months: a psychometric study, Res. Dev. Disabil. 110 (2021), 103837.
[13] E.S.L. dos Santos, J.F. de Kieviet, M. Königs, R.M. van Elburg, J. Oosterlaan, Predictive value of the Bayley scales of infant development on development of very preterm/very low birth weight children: a meta-analysis, Early Hum. Dev. 89 (7) (2013) 487–496.
[14] D.E. Creighton, S. Tang, J. Newman, L. Hendson, R. Sauve, Establishing Bayley-III cut-off scores at 21 months for predicting low IQ scores at 3 years of age in a preterm cohort, Paediatr. Child Health 23 (8) (2018) e163–e169.
[15] M.M. Bode, D.B. D'Eugenio, B.B. Mettelman, S.J. Gross, Predictive validity of the Bayley, at 2 years for intelligence quotient at 4 years in preterm infants, J.Dev.Behav.Pediatr. 35 (9) (2014) 570–575.
[16] J. Månsson, K. Stjernqvist, F. Serenius, U. Ådén, K. Källén, Agreement between Bayley-III measurements and WISC-IV measurements in typically developing children, J. Psychoeduc. Assess. 37 (5) (2019) 603–616.
[17] J.D. Hamadani, H. Baker-Henningham, F. Tofail, F. Mehrin, S.N. Huda, S. M. Grantham-McGregor, Validity and reliability of mothers' reports of language development in 1-year-old children in a large-scale survey in Bangladesh, Food Nutr. Bull. 31 (2_suppl2) (2010) S198–S206.
[18] S.E. Fox, P. Levitt, C.A. Nelson 3rd., How the timing and quality of early experiences influence the development of brain architecture, Child Dev. 81 (1) (2010) 28–40.
[19] L.C. Fernald, E. Prado, P. Kariger, A. Raikes, A Toolkit for Measuring Early Childhood Development in Low and Middle-income Countries, 2017.
[20] T.A. Strand, M. Ulak, R.K. Chandyo, I. Kvestad, M. Hysing, M. Shrestha, et al., The effect of vitamin B12 supplementation in Nepalese infants on growth and development: study protocol for a randomized controlled trial, Trials 18 (1) (2017) 187.
[21] T.A. Strand, M. Ulak, M. Hysing, S. Ranjitkar, I. Kvestad, M. Shrestha, et al., Effects of vitamin B12 supplementation on neurodevelopment and growth in Nepalese infants: a randomized controlled trial, PLoS Med. 17 (12) (2020), e1003430.
[22] S.R. Psaki, J.C. Seidman, M. Miller, M. Gottlieb, Z.A. Bhutta, T. Ahmed, et al., Measuring socioeconomic status in multicountry studies: results from the eight-country MAL-ED study, Popul. Health Metrics 12 (1) (2014) 1–11.
[23] N. Bayley, Bayley Scales of Infant and Toddler Development, PsychCorp, Pearson, 2006.
[24] D. Wechsler, Wechsler preschool and primary scale of intelligence -, Fourth Edition, Pearson Assessment, London, 2012.
[25] L. Lin, L.D. Torbeck, Coefficient of accuracy and concordance correlation coefficient: new statistics for methods comparison, PDA J. Pharm. Sci. Technol. 52 (2) (1998) 55–59.
[26] M. Augustyniak, D. Mrozek-Budzyn, A. Kieltyka, R. Majewska, Stability of the mental and motor Bayley scales of infant development in infants over first three years of life, Przegl. Epidemiol. 67 (3) (2013) 483–486.
[27] L. Schonhaut, M. Pérez, I. Armijo, A. Maturana, Comparison between Ages & Stages Questionnaire and Bayley Scales, to predict cognitive delay in school age, Early Hum. Dev. 141 (2020), 104933.

[28] S. Ranjitkar, M. Hysing, I. Kvestad, M. Shrestha, M. Ulak, J.S. Shilpakar, et al., Determinants of cognitive development in the early life of children in BhaktapurNepal, Front. Psychology 10 (2019) 2739.
[29] B.J. McCormick, L.E. Caulfield, S.A. Richard, L. Pendergast, J.C. Seidman, A. Maphula, et al., Early life experiences and trajectories of cognitive development, Pediatrics 146 (3) (2020).
[30] D.C. McCoy, E.D. Peet, M. Ezzati, G. Danaei, M.M. Black, C.R. Sudfeld, et al., Early childhood developmental status in low- and middle-income countries: national, regional, and global prevalence estimates using predictive modeling, PLoS Med. 13 (6) (2016), e1002034.