



Validity and cross-cultural differences of the Bayley Scales of Infant and Toddler Development, Third Edition in typically developing infants

Jasmine Hoskens^{a,b,*}, Katrijn Klingels^{a,b,1}, Bouwien Smits-Engelsman^{c,1}

^a KU Leuven - University of Leuven, Faculty of Kinesiology and Rehabilitation Sciences, Department of Rehabilitation Sciences, Neuromotor Rehabilitation Research Group, Herestraat 49, Box 1510, B-3000 Leuven, Belgium

^b Hasselt University, REVAL Rehabilitation Research Center, BIOMED, Agoralaan gebouw D, B-3590 Diepenbeek, Belgium

^c Department of Health and Rehabilitation, University of Cape Town, Old Main Building, Groote Schuur Hospital, Cape Town, South Africa

ARTICLE INFO

Keywords:

Developmental assessment
Motor development
Bayley-III
Validity
Cross-cultural differences

ABSTRACT

Background: The Dutch translation of Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III-NL) is a comprehensive tool assessing cognitive, language and motor development in children up to 42 months.

Aims: The first aim of this study was to evaluate concurrent validity of the Bayley-III-NL Gross Motor Scale (GMS) in relation to the Alberta Infant Motor Scale (AIMS). Secondly, divergent validity between the other subscales mutually and the AIMS was investigated. Finally, the importance of population-specific reference values was examined.

Methods and procedures: A sample of 122 healthy, typically developing Flemish children (mean age: 9 months 7 days), born full term was assessed with the Bayley-III-NL and the AIMS.

Outcomes and results: Concurrent validity of the Bayley-III-NL GMS and the AIMS was moderate to high (0.59–0.98; $p < 0.001$). In addition, weak correlations (-0.10 – 0.27) between the non-motor-Bayley-III-NL subscales and the AIMS were found. Finally, significant differences were found between Bayley-III scores based on Flemish and American norms ($p < 0.001$), except for fine motor skills ($p = 0.11$).

Conclusions and Implications: This study provides support for the concurrent validity of the Bayley-III-NL GMS and divergent validity of the different Bayley-III-NL subscales. Secondly, population-specific reference values should be used to avoid over- and under estimation of infant's development.

1. Introduction

During the first years of life, children develop on different developmental domains and at varying rates in a very complex manner. Motor development is an important manifestation of the integrity and functionality of the central nervous system. Early detection of deviations in motor development enables health professionals to diagnose specific disabilities and if required, to start targeted interventions. This early detection is crucial to reach optimal outcome with early intervention, as plasticity of the developing brain can still occur. Thus, there is a strong need for reliable, valid and norm referenced tests to assess early motor development.

The Bayley Scales of Infant and Toddler Development-Third Edition (Bayley-III) and the Alberta Infant Motor Scale (AIMS) are two commonly used tools to assess motor development in young infants [1,2].

The Bayley-III is an individually administered instrument assessing development of infants and young children between 16 days and 42 months of age [1]. Its primary purposes are to identify children with developmental delay and to provide information for intervention planning. The instrument was updated from the Bayley Scales of Infant Development- 2nd edition (BSID-II) in the following aspects: revision and construction of test items, simplification of the assessment procedure, updating of the stimulus materials and standardization of a new population. The original Bayley-III normative data were established in 1700 American infants [1] and in 1192 Flemish children [3] divided over 17 age groups. The third version of the test consists of five distinct scales: cognition, language, motor skills, social-emotional and adaptive behavior. The language scale is subdivided into receptive and expressive communication, whereas the motor scale contains a fine motor and gross motor scale.

* Corresponding author at: KU Leuven - University of Leuven, Faculty of Kinesiology and Rehabilitation Sciences, Department of Rehabilitation Sciences, Neuromotor Rehabilitation Research Group, Herestraat 49, Box 1510, B-3000 Leuven, Belgium.

E-mail addresses: jasmine.hoskens@kuleuven.be (J. Hoskens), katrijn.klingels@Uhasselt.be (K. Klingels).

¹ These authors have contributed equally to this work.

The Alberta Infant Motor Scale (AIMS) is a Canadian norm-referenced measure to assess the gross motor abilities of infants from birth to independent walking [2]. The AIMS focuses on the attainment of motor milestones in four positions: prone, supine, sitting and standing. The AIMS normative data comprise 2202 Canadian infants between 1 week and 18 months old, born between 1990 and 1992 [2].

Concurrent validity of the Bayley-III and AIMS has recently been reported in preterm infants and children with risk at or with a developmental delay, but no studies are yet available in typically developing children [4,5]. Concurrent validity of the Bayley-III-NL is described in the manual, but also for this version no comparison has been made yet with another motor test. Earlier studies indicate the need to determine whether revisions made on the Bayley scales strengthen the concurrent validity with other tools used for clinical identification of children with developmental delays [6].

When a scale is used in a different population, it is also necessary to redefine its reference values to overcome over and under referral of children. Different cross-sectional studies in Western countries indicate that the Bayley-III tends to overestimate the performance of children at risk for developmental disorders such as prematurity, autism spectrum disorders and heart disease [7–11]. A study of Yu et al. also found underestimation of the rates of developmental delays in both term and preterm Taiwanese infants by the Bayley-III [12]. A recent study of Steenis et al. showed that using the US instead of Dutch norms resulted in over-referral regarding gross motor skills, and under-referral regarding cognitive, receptive and expressive communication, and fine motor skills [13]. Also the Bayley-III norms developed for Malawian children were not equivalent to the US-based norms, but differed across age groups and also between subtests [14].

To date, Bayley-III-NL results collected in Flanders have not yet been compared to the American norm-referenced values. On the other hand, Fleuren et al. and De Kegel et al. compared the AIMS results of Dutch and Flemish children respectively to its Canadian norm-referenced values. The children in both studies showed significantly lower motor scores than the Canadian norm population [15,16].

Therefore, the purpose of this study was to provide further psychometric evidence for the Bayley-III-NL in a large sample of young typically developing infants recruited in Flanders, Belgium. The aim of this study was threefold. The first aim was to evaluate the concurrent validity of the Bayley-III-NL GMS and the AIMS. In addition, the subscales of the AIMS in the four different positions (prone, supine, sitting and standing) were compared to specifically selected items of the Bayley-III-NL GMS in the corresponding positions. Secondly, divergent validity was evaluated by comparing the other (non-motor) Bayley-III-NL scales with the AIMS and with each other. The last objective was to investigate cross-cultural differences in development of young children and subsequently the need for population-specific norm values.

2. Materials and methods

2.1. Participants

This study included a sample of 122 Flemish children, aged between 16 days and 18 months old. All infants were healthy, typically developing and born full-term. Exclusion criteria were prematurity or known medical problems that could have influenced development, such as sensory problems (e.g. deafness, blindness), motor difficulties, speech or language deficits, or learning and behavioral problems. A medical health questionnaire for the parents was used to ascertain these criteria. Children were randomly recruited via childcare centers, home care providers or by direct contact with parents. All parents who were willing to take part signed a written consent form prior to participation in accordance with the declaration of Helsinki. The study (S53591) was approved by the local ethics committee (Commissie Medische Ethiek KU Leuven).

2.2. Instruments and procedures

Child development was assessed with the Bayley-III-NL, the Dutch translation of the Bayley-III, which contains five scales: cognition (91 items), receptive (49 items) and expressive communication (46 items) and fine (66 items) and gross motor function (72 items) [1,3].

The Dutch translation of the Bayley-III (Bayley-III-NL) has shown good test-retest reliability ($r = 0.72$ – 0.78) and inter-rater reliability (average kappa = 0.77) [3]. Only low to moderate correlations were found between the Bayley-III-NL and BSID-II [3]. Moderate to high concurrent validity has been shown for the Bayley-III-NL cognition and language scales and the Dutch version of the Wechsler Preschool and Primary Scale of Intelligence, third Edition (WPPSI-III-NL), the Lexilijst and Schlichting test for receptive and expressive communication [3]. Bayley [1] found moderate to high correlations between the Bayley Scales of Infant Development-Second edition (BSID-II) and Peabody Developmental Motor Scales, second edition (PDMS-II) [1]. Despite similarities in the language between Flanders and the Netherlands, differences found during the process of collecting norm data led to the conclusion that it was necessary to develop separate norms for both populations.

The Bayley-III-NL assesses the level of the child's development in a playful way using engaging toys and activities. The total raw score of each scale was calculated as the sum of the items for which the child received a credit along with the number of unadministered items preceding the basal level. Besides raw scores, the Bayley-III provides also four types of norm-referenced scores. For the subscales we used scaled scores (mean = 10; SD = 3). Subsequently percentile ranks were calculated for the overall cognition, communication and motor subdomains. Norm-referenced scores were calculated both based on the Flemish reference values as well as based on the American reference values.

Children were also evaluated with the Alberta Infant Motor Scale (AIMS) by the same examiner. The AIMS consists of 58 gross motor items, organized into four subscales: prone (21 items), supine (9 items), sitting (12 items) and standing (16 items) [2]. The AIMS has shown to be a reliable and valid measure of motor development for infants at risk for motor delay [2]. The concurrent validity was established by high correlations between AIMS scores and scores on the PDMS and the BSID-II [2].

The AIMS observes spontaneous behavior in four different positions without specific equipment. The score of each subscale is calculated as the sum of all observed items, plus all items prior to the first observed item. The total AIMS score is the sum of the four subscale scores. The raw AIMS scores were also converted to age-based percentile ranks.

All children were evaluated by three physiotherapists. Prior to the assessments, these testers underwent formal training, starting with a two-day education on administering the Bayley-III-NL, followed by three theoretical cases and five test-videos to optimize standardization and inter-rater reliability. The three theoretical cases were provided on paper as an exercise for scoring and interpretation. Subsequently, they had to perform and score five assessments, which were video-taped. The videos were evaluated by the trainers on standardization of assessment, scoring and interpretation. The evaluators were also trained in scoring the AIMS. All children were first assessed with the Bayley-III-NL and then with the AIMS by the same tester during one assessment. The assessments were performed at day care or in a home setting. Administration time ranged from 30 to 90 min, depending on the child's age. The most optimal test environment (quiet, good lighting, minimal distractions) was ensured and evaluation took only place when the infant was alert and cooperative. A caregiver was allowed to be present besides the child and the examiner.

2.3. Statistical analysis

Descriptive statistics (mean, standard deviation, range) were used to

document subject characteristics. Normality was checked using the Shapiro Wilk test and visual inspection of the data. Parametric test statistics were used. To evaluate concurrent validity, Pearson product-moment correlation coefficients were calculated between the Bayley-III-NL motor percentile and the AIMS percentile and between the total raw score of the Bayley-III-NL Gross Motor Scale (GMS) scores with the total AIMS score. Subsequently, selected items of the Bayley-III-NL GMS in the four corresponding positions of the AIMS subscales were compared to the four AIMS subscales. For this purpose, observational items up to and including item 46 ('Stands up, Advanced') were extracted from the Bayley-III-NL GMS, and classified into the four positions: supine (6 items), prone (11 items), sitting (6 items) and standing (13 items). The selected items per position are described in [Appendix A](#). A scatter plot with linear regression line and 95% confidence intervals was made to illustrate the relations between individual results on the AIMS and Bayley-III-NL GMS.

Secondly, to investigate divergent validity Pearson product moment-correlations between the other Bayley-III-NL scales and the AIMS and between the different Bayley-III-NL scales mutually were calculated.

Correlation coefficients of > 0.70 were considered as high, 0.50 – 0.70 as moderate, 0.30 – 0.50 as fair and < 0.30 as weak or no association [17].

Finally, a comparison was made between the Bayley-III scaled scores of the Flemish sample according to the Flemish and the corresponding American reference values. Flemish and American/Canadian mean, standard deviation and range of the scaled scores and motor percentiles of the Bayley-III and AIMS were calculated. The Flemish Bayley-III motor percentile was thereafter compared to the Canadian AIMS percentile. To check if differences between groups were age-dependent, the infants were also divided in 6 different age groups: 15 days–3 months 15 days; 3 months 16 days–6 months 15 days; 6 months 16 days–9 months 15 days; 9 months 16 days–12 months 15 days; 12 months 16 days–15 months 15 days; 15 months 16 days–18 months 15 days. A paired *t*-test was used to compare means and Bonferroni-Holm Correction was used as correction for multiple comparisons.

Lastly, the motor percentiles according to the Bayley-III-NL (Flemish) and AIMS were compared to the cut-off values used in clinical practice (percentiles 16 and 5).

Data analyses were done using SPSS20; *p*-values below 0.05 were considered significant.

3. Results

A sample of 122 children (62 boys; 60 girls) was included. The mean age was 9 months 7 days with a standard deviation of 5 months 5 days (range: 19 days–18 months 1 day). The mean birth weight was 3384 g (SD = 435 g and range: 2050–4330 g). All children were Caucasian and born in Belgium. The overall education level of the parents was high, 88% of the mothers and 73% of the fathers went to higher professional education or university. [Table 1](#) gives an overview of the population and background characteristics. [Table 2](#) presents the means, standard deviations and ranges of the total raw scores, scaled scores and percentiles on the Bayley-III-NL subscales and the AIMS.

A moderate correlation was found between the Bayley-III-NL motor percentile (including both fine and gross motor subscales) and the AIMS percentile ($r = 0.59$; $p < 0.001$), whereas a very high correlation between the Bayley-III GMS raw scores and the AIMS total raw scores was found ($r = 0.98$; $p < 0.001$). Correlation coefficients of the Bayley-III-NL GMS raw scores (in supine, prone, sitting and standing) and the AIMS subscales scores in the corresponding positions were also very high, ranging between 0.89 and 0.96. All correlations are listed in [Table 3](#).

[Figure 1](#) shows the linear relationship between the Bayley-III-NL GMS and the AIMS total scores in a scatter plot with regression line and

Table 1
Population and background characteristics.

Characteristics	n	
Age	122	Mean (SD): 9 mo 7 d (5 mo 5 d) Range: 19 d – 18 mo 1 d
Gender	122	Male: 62; female 60
Gestational age	121	Mean (SD): 39,4 w (1.2 w) Range: 37 w – 42 w
Birth Weight	120	Mean (SD): 3384 g (435 g) Range: 2050 g–4330 g
Ethnicity parents (other country of birth)	121	Flemish: 117; non-Belgian: 4 (1 Bulgaria; 2 Poland; 1 Russia)
Education level mother	120	Secondary education: 14 Higher professional or university: 106
Education level father	119	Secondary education: 32 Higher professional or university: 87

n: number; SD: standard deviation; mo: months; d: days; w: weeks; g: grams.

95% confidence interval. The lowest and highest scores on the AIMS form an approximately horizontal line, meaning that in these children, there is more variation within the scores of Bayley-III-NL GMS compared to the AIMS score. For the same AIMS score different Bayley-III-NL scores were found.

Secondly, a fair correlation was found between the Bayley-III-NL Fine Motor (FM) scaled scores and the GM scaled scores ($r = 0.40$; $p < 0.001$). Low correlation coefficients ($r = -0.10$ – 0.27) were found between the cognition, language and motor percentiles of the Bayley-III-NL and the AIMS as well as between these different Bayley-III-NL percentiles ([Table 3](#)).

Finally, Bayley-III scaled scores and percentiles according to Flemish and American reference values were compared. [Table 4](#) shows mean, standard deviation and range of the scaled scores of the different Bayley-III scales and the motor percentiles according to Bayley-III-NL and AIMS. The scaled scores of the total group according to the American reference values were significantly lower than the scores according to the population-specific Flemish reference values for all Bayley-III subscales ($p < 0.001$), except for the fine motor scale ($p = 0.11$) ([Table 4](#)). Similarly, children of all age groups scored significantly lower on the two communication scales according to the American reference values. For the cognition scale, the American scaled scores were overall lower as well, except for the two oldest age groups where the Flemish scaled scores were significantly higher. The youngest children between 15 days and 3 months 15 days scored significantly higher based on the American reference values on the gross motor scale, while for all other age groups significantly lower scores were found. Regarding fine motor skills scales, no significant differences were found between scaled scores based on Flemish or American reference values ([Table 4bis](#)).

For the total group, a significant difference was found between the Bayley-III-NL motor percentile based on the Flemish reference values (mean = 47.12; SD = 27.85) and the Canadian-based AIMS percentile (mean = 26.86; SD = 27.15) ($p < 0.001$). The difference in motor percentiles was not significant for the youngest and oldest age groups ($p = 0.13$ and $p = 0.85$, respectively). Only 16 out of the 122 children scored below the 16th percentile (P16) and only 5 below P5 regarding the Bayley-III-NL motor percentile, while 60 children scored below P16 and even 27 below P5 on the AIMS (Canadian reference values).

4. Discussion

The purpose of this study was to provide further psychometric evidence for the Bayley-III-NL on a large sample of young typically developing infants recruited in Flanders, Belgium by investigating concurrent and divergent validity of the Bayley-III-NL with the AIMS and by exploring the relation between the different early developmental domains. The need for population-specific reference values was

Table 2

Mean total raw scores, scaled scores and percentiles with standard deviations and ranges of the Bayley-III-NL subscales and the AIMS.

Scoring (n = 122)	Total raw scores			Scaled scores			Percentiles		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Bayley-III-NL scales									
Cognition (0–91)	31.10	14.78	3–54	10.20	2.35	3–15	52.17	Cognition 25.38	1–95
Receptive communication (0–49)	11.19	4.32	3–21	11.31	2.63	5–17		Communication 25.89	2–99
Expressive communication (0–46)	10.30	5.50	1–24	11.34	2.77	4–18	65.02		
Fine motor (0–66)	22.09	10.28	2–40	9.98	2.51	3–17		Motor 27.85	1–96
Gross motor (0–72)	28.25	14.21	2–50	9.48	2.82	4–16	47.12		
AIMS (0–58)	32.34	19.51	3–58	/	/	/	26.86	27.15	0.5–90

Bayley-III-NL: The Dutch translation of the Bayley Scales of Toddler and Infant Development, Third Edition; n: number; SD: standard deviation; scaled scores: mean = 10; SD = 3; AIMS: Alberta Infant Motor Scale.

Table 3

Concurrent and divergent validity Pearson product-moment correlation coefficients.

	Pearson's r	p-Value
Concurrent validity		
Bayley-III-NL motor percentile - AIMS percentile	0.59	< 0.001*
Bayley-III-NL GMS - AIMS total raw scores	0.98	< 0.001*
Bayley-III-NL GMS supine - AIMS supine raw scores	0.89	< 0.001*
Bayley-III-NL GMS prone - AIMS prone raw scores	0.93	< 0.001*
Bayley-III-NL GMS sitting - AIMS sitting raw scores	0.96	< 0.001*
Bayley-III-NL GMS standing - AIMS standing raw scores	0.96	< 0.001*
Divergent validity		
Bayley-III-NL FMS scaled score - Bayley-III GMS scaled score	0.40	< 0.001*
Bayley-III-NL cognition percentile - AIMS percentile	−0.10	0.27
Bayley-III-NL language percentile - AIMS percentile	0.12	0.18
Bayley-III-NL cognition percentile - Bayley-III language percentile	0.27	0.003*
Bayley-III-NL cognition percentile - Bayley-III motor percentile	0.19	0.04*
Bayley-III-NL language percentile - Bayley-III motor percentile	0.13	0.15

Pearson's r: Pearson product-moment correlation coefficient; Bayley-III-NL: The Dutch translation of the Bayley Scales of Toddler and Infant Development, Third Edition; AIMS: Alberta Infant Motor Scale; GMS: gross motor scale (total raw score); FM: fine motor; GM: gross motor.

* p < 0.05.

also investigated by comparing the results of the Flemish children to the American and Canadian reference populations of the Bayley-III and AIMS, respectively.

The concurrent validity of the Bayley-III-NL GMS and the AIMS was very high with an overall correlation coefficient of 0.98 and correlation coefficients above 0.89 for the comparison between the different positions (prone, supine, sitting and standing). Visualization of the individual scores showed that the Bayley-III-NL GMS discriminated better than the AIMS in children at the lower and upper end of the AIMS. This might be because the AIMS only observes spontaneous behavior, whereas the Bayley-III-NL utilizes external feedback and equipment. Besides the selected items of the Bayley-III-NL GMS in the four positions of the AIMS also more specific observations of e.g. the position of the head during carrying, compensations of the head during tilting or throwing a ball, are evaluated. Although in general, the items of the Bayley-III-NL GMS and the AIMS are very similar, a small difference is that in the AIMS the items are chronologically grouped per position, while the Bayley-III-NL evaluates the items in a chronological order but not grouped per position. Moreover, the Bayley-III-NL has a wider age range. Most children will reach the ceiling score on the AIMS even already just before 18 months. For those children the Bayley-III-NL still has more items to make a better and more accurate estimation of the motor performance.

The results of our study confirm the results of other studies, investigating the correlation between the AIMS and the BSID-II motor scale in preterm infants or high-risk infants [18–20]. Worth noting is that the BSID-II Motor Scale consists of a combination of fine and gross motor skills, while separate subscales are defined in the Bayley-III. Recently, De Albuquerque et al. investigated the concurrent validity of the AIMS in 159 preterm infants, using linear regression and ceiling effect analysis. A determination coefficient (R^2) of 0.92 was found, indicating a good predictive validity of the AIMS and a good correlation with the Bayley-III GMS [21]. Subsequently, a Spanish version of the AIMS was also compared to the Bayley-III GMS in a sample of 50 infants at risk for developing or had already developed a motor delay or disorder [5]. Moderate to very high correlation coefficients were found for different age groups, ranging from 0.69 to 0.97 [5]. Notwithstanding that, previous studies used older versions of the Bayley Scales or investigated preterm children or children at risk for or with motor delays, our study supports the good concurrent validity of the Bayley-III with the AIMS in typically developing children.

Our second research question investigated the divergent validity between the Bayley-III-NL cognition, communication (receptive and expressive) and motor (fine and gross) scales mutually and the AIMS, proving that the different subscales measure different skills. Weak correlations were found between the different subscales of the Bayley-III-NL and the AIMS. Campos et al. evaluated the motor and cognitive performance in 94 infants with the BSID-II at 1, 2, 3, 6, 9 and 12 months [23]. The results showed significant differences between motor and cognitive performance at 1, 2, and 3 months of age. However, at 6, 9, and 12 months, there was no difference between these domains [23]. It should be noted that in the BSID-II gross and fine motor development are evaluated as one (motor) domain and in that version several fine motor and language skills were embedded in the cognition scale. In another study of Souza et al. the gross and fine motor skills of infants were compared in 30 infants with the Bayley-III once at 12 months and a second time at 17 months. They found a large discrepancy between the two subscales of the Bayley-III. Likewise, we only found a fair correlation between the gross and fine motor skills of the Bayley-III-NL. The results of our study and previous studies prove that the Bayley-III is able to measure the different developmental domains separately and specifically.

Finally, a comparison was made between the Bayley-III scores according to the Flemish and American norm values and according to the Canadian norm values of the AIMS. Significant lower scores were found according to the American norm values on the receptive and expressive communication and gross motor scale, indicating that Flemish children perform less well on those subscales of the Bayley-III. Therefore, using the American norm values instead of the Flemish norm values in Flemish children will result in over-referral. The largest differences between Flemish and American-based norm scores were seen for the Bayley-III GMS, in children between 6 months 16 days and 15 months 15 days. For those age groups the differences between both scores are

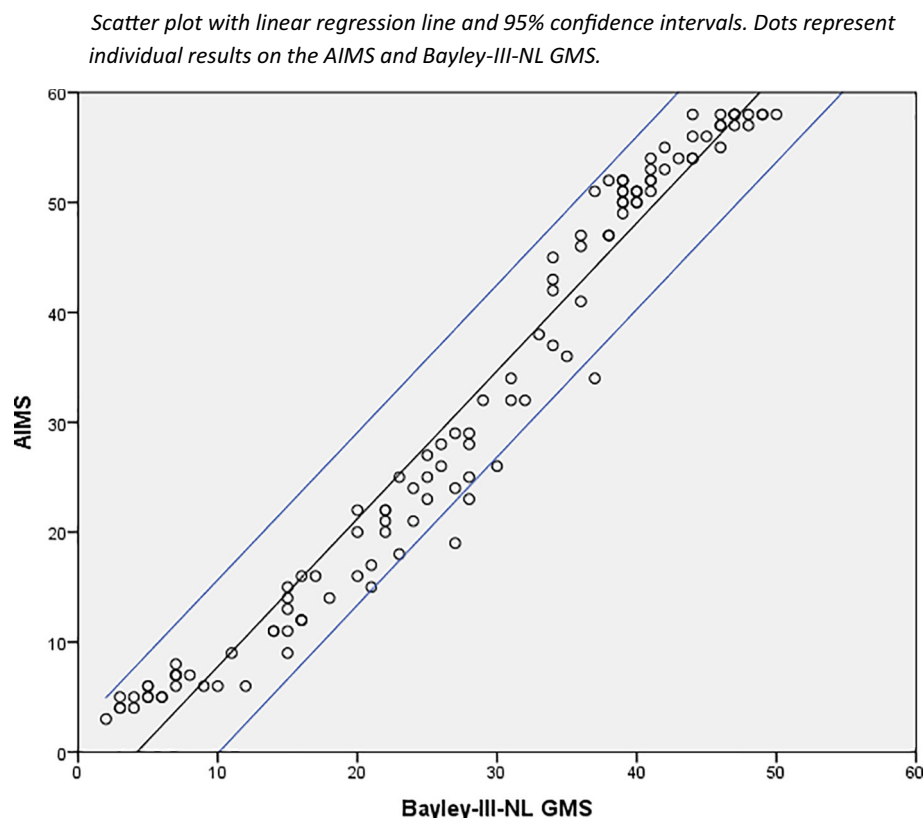


Figure 1. AIMS: Alberta Infant Motor Scale; Bayley-III-NL GMS: The Dutch translation of the Bayley Scales of Toddler and Infant Development, Third Edition Gross Motor Scale.

more or less one standard deviation, which might be clinically relevant. These findings correspond with the results of [13], who found the largest differences between Dutch and American scores also on GMS of the Bayley-III-NL in children of approximately nine months old [13].

Concerning the cognitive domain, results varied more between age groups. No significant differences were found for the fine motor scales. Other studies also found different scores using the population-specific norms instead of the American norm values of the Bayley-III. Cromwell

et al. showed that the Bayley-III norms developed for healthy Malawian children were not equivalent to the US-based norms. The relationship between the two norms was not uniform but differed across age groups and differed for each subtest [14]. Steenis et al. reported similar results in the Netherlands with significant differences in means across all five subscales using American and Dutch norm values [13]. The overall results differ from our results as they found differences in means across all five subtests ($p < 0.01$) with small to large effect sizes (ranging

Table 4

Comparison between mean scaled scores of the Bayley-III according to the Flemish and American norm values and between mean motor percentiles of the Bayley-III Motor Scale and Alberta Infant Motor Scale (according to Flemish and Canadian norm values).

Bayley-III (scaled scores)	Flemish			American			p-Value
	Mean	SD	Range	Mean	SD	Range	
n = 122							
Cognition scale	10.20	2.35	3–15	9.57	2.01	1–14	< 0.001*
Receptive communication scale	11.31	2.63	5–17	9.02	2.40	4–15	< 0.001*
Expressive communication scale	11.34	2.77	4–18	8.88	2.22	4–15	< 0.001*
Fine motor scale	9.98	2.51	3–17	10.14	2.00	5–16	0.11
Gross motor scale	9.48	2.82	4–16	7.91	2.15	3–13	< 0.001*

Motor percentiles	Flemish			Canadian			p-Value
	Bayley Motor Percentile			AIMS Total Percentile			
n = 122							
	47.12	27.85	1–96	26.86	27.15	0.5–90	< 0.001*

Bayley-III: Bayley Scales of Toddler and Infant Development, Third Edition; n: number; SD: standard deviation; scaled scores: mean = 10; SD = 3; AIMS: Alberta Infant Motor Scale.

* Significant after Bonferroni-Holm Correction ($p = 0.05$).

Table 4bis

Comparison between mean scaled scores of the Bayley-III according to the Flemish and American norm values and between mean motor percentiles of the Bayley-III Motor Scale and Alberta Infant Motor Scale (according to Flemish and Canadian norm values) per age group.

Age group	n	Scales	Flemish			American			P-value
(months: days)			Mean	SD	Range	Mean	SD	Range	
00:15- 03:15	23	CS	9.30	1.80	6-14	8.39	1.16	5-11	0.012
		RCS	10.91	2.15	6-17	9.61	1.70	6-12	<0.001*
		ECS	10.65	2.57	5-16	8.57	1.56	5-11	<0.001*
		FMS	8.78	2.78	3-13	9.78	1.93	5-13	0.007
		GMS	7.17	2.10	4-12	8.96	1.43	7-12	<0.001*
			Flemish			Canadian			
			Bayley Motor Percentile			AIMS Total Percentile			P-value
			26.70	21.44	1-75	20.91	19.38	2-64	0.13
03:16- 06:15	24	CS	12.33	1.99	5-15	9.79	2.48	1-14	<0.001*
		RCS	11.67	2.58	6-17	9.75	2.66	4-15	<0.001*
		ECS	11.96	2.51	8-18	9.54	2.34	6-15	<0.001*
		FMS	9.50	1.87	5-12	9.38	1.86	6-12	0.33
		GMS	9.46	2.15	6-14	8.25	2.21	4-13	<0.001*
			Bayley Motor Percentile			AIMS Total Percentile			P-value
			42.96	23.12	8-84	17.60	17.79	0.5-65	<0.001*
06:16- 09:15	13	CS	11.62	2.22	8-15	9.85	2.34	5-13	<0.001*
		RCS	11.77	2.80	8-17	8.85	2.85	5-14	<0.001*
		ECS	10.69	3.77	6-16	8.31	3.20	5-13	<0.001*
		FMS	11.08	2.75	8-16	11.38	2.60	8-16	0.04
		GMS	10.46	2.33	7-16	7.23	2.13	4-12	<0.001*
			Bayley Motor Percentile			AIMS Total Percentile			P-value
			59.15	25.67	21-95	11.65	10.53	0.5-32	<0.001*
09:16- 12:15	25	CS	9.88	1.83	7-13	9.80	1.78	7-13	0.16
		RCS	11.80	2.47	7-17	8.60	2.40	4-14	<0.001*
		ECS	12.04	2.42	6-17	9.08	2.18	4-13	<0.001*
		FMS	10.12	1.94	6-13	10.16	1.34	8-13	0.83
		GMS	10.48	3.57	5-15	7.16	2.72	3-11	<0.001*
			Bayley Motor Percentile			AIMS Total Percentile			P-value
			53.72	31.62	5-95	27.70	25.56	0.5-73	<0.001*
12:16- 15:15	22	CS	9.18	1.99	6-13	9.95	1.56	7-13	<0.001*
		RCS	11.23	2.84	7-17	8.68	2.38	5-14	<0.001*
		ECS	11.77	2.49	9-18	8.95	1.94	6-14	<0.001*
		FMS	9.91	2.27	5-15	9.82	1.71	7-14	0.61
		GMS	10.18	2.72	5-16	7.77	2.02	4-13	<0.001*
			Bayley Motor Percentile			AIMS Total Percentile			P-value
			50.91	25.02	2-96	28.84	30.12	0.5-90	0.001*
15:16- 18:15	15	CS	8.93	2.40	3-12	9.80	2.08	5-13	<0.001*
		RCS	10.27	3.15	5-15	8.27	2.38	4-11	<0.001*
		ECS	10.20	3.08	4-14	8.33	2.38	4-11	<0.001*
		FMS	11.53	3.07	6-17	11.27	2.34	7-16	0.22
		GMS	9.47	2.13	5-13	7.80	1.57	4-10	<0.001*
			Bayley Motor Percentile			AIMS Total Percentile			P-value
			58.13	28.34	5-92	59.67	33.42	1-90	0.85

Bayley-III: Bayley Scales of Toddler and Infant Development, Third Edition; n: number; SD: standard deviation; CS: cognition scale; RCS: receptive communication scale; ECS: expressive communication scale; FMS: fine motor scale; GMS: gross motor scale; AIMS: Alberta Infant Motor Scale; *Significant after Bonferroni-Holm Correction ($p = 0.05$).

from 0.03 to 0.26) [13]. Significantly lower American scores were found on the gross motor scale, resulting in over-referral, whereas under-referral would be seen for the other four subscales using the American norm values, although the results regarding over- and under-referral are to some extent also age-dependent [13]. Remarkably, even differences exist in reference values between Flanders and the Netherlands, two neighboring countries with a comparable culture and population. A possible explanation might be the difference in childcare. An important difference is that in the Netherlands the 0–2 year old children spend on average only 17–18 h a week in formal care, whereas their Belgian peers spend between 25 and 35 h a week in daycare [24]. In general, it is hypothesized that more hours in childcare are associated with better language abilities in early childhood. However a study of [25] in the Netherlands showed that more hours in care in the first year of life were associated with lower levels of language proficiency at ages 1 to 1.5 [25]. At later ages, this effect disappeared and language proficiency increased [25]. These findings were consistent with the results of an American and Norwegian study, demonstrating that childcare experience before the age of one is associated with lower early languages scores, whereas childcare after the age of 1 is associated with higher language scores [26,27]. The Dutch people are also known as more abundant and direct in communication than Flemish people. This might have influenced the communication scales. Importantly, there is not only a wide variation in childcare usage between countries, but another important factor related to this is the socio-economic background of the families [28].

The Flemish Bayley-III-NL motor percentile was also compared to the total AIMS percentile. For the total group and all subgroups except the youngest and the oldest, we found significantly lower AIMS percentiles compared to the Bayley-III-NL motor percentiles. These results corroborate with the results of Fleuren et al. and De Kegel et al., who found significantly lower motor scores in Dutch and Flemish children on the AIMS compared to the Canadian reference population [15,16]. In Belgium, the 16th percentile (P16) is used as overall cut-off score to start motor therapy. In our study 60 children (49%) scored below P16 and even 27 children (22%) scored below P5 on the AIMS. This means that 27 of the presumably typically developing children scored almost two standard deviations below the mean, whereas regarding the Bayley-III-NL only five children scored this low on the motor domains. By relying on the Canadian norms of the AIMS an over-referral of children will occur. On the other hand, Darrah et al. investigated in 2014 if the infant gross motor abilities changed in 20 years, by a re-evaluation of the normative values in Canada. They showed that the sequence and age at emergence of the AIMS items has remained similar over 20 years and that therefore the current normative values remain valid [29].

The significant differences between Flemish, American and Canadian based scores support the need for population specific reference values to prevent misclassification of our children. To date, we can only hypothesize about the reasons for the differences in scores between Flemish, American and other children. There might be differences in parental information provided by the government. For example, in Flanders in 1995 the 'Back to sleep' campaign to prevent Sudden Infant Death was introduced. This campaign is still supported by childcare offices and has some implications regarding motor development. The study of [15] proved that the majority of our Flemish children slept in supine during the first 6 months of life and that they also rarely play in prone during the day. The Flemish campaign is primarily concentrated on the supine sleep position and the sleep environment while promotion of tummy time is barely cited. As promoted by the American Academy of Pediatrics in 2008, the information for

parents must change into 'Back to Sleep, Tummy to Play' [30]. Variation in positions and environment is crucial for early motor development. Prone lying is for example important to develop upper body strength in order to overcome gravity and so optimize early motor milestone acquisition [31]. Several studies have shown that children positioned in supine for sleeping, reach motor milestones at a later age [31–34]. Another possible explanation might be the increasing use of equipment to position children. Nowadays, children are placed more frequently and at an earlier age in different supportive seating devices and baby walkers, even before they have reached the related milestones compared to children in the early nineties. This gives them fewer opportunities to move around and explore the environment, which might have a negative influence on early gross motor development and explain the lower scores using the Canadian reference values of the AIMS.

In short, many cultural and region dependent factors such as childcare, education, and socio-economic status have an influence on child development. Therefore, we want to emphasize the importance of country- or region-specific reference values to evaluate early development, which is also pledged by other authors [13,15,16,35].

Some limitations of this study should be considered. First, all children were assessed by one of three different evaluators. Furthermore, the Bayley-III-NL and the AIMS were assessed during the same session, by the same evaluator. Administration of both scales by the same evaluator might have resulted in assessment bias and the duration was fairly long which might have influenced the children's performance. After the evaluation with Bayley-III-NL, the children had to be placed again in supine, prone, sitting and standing position to complete the scoring of the AIMS, because some items differ from the Bayley-III-NL or require different scoring criteria. However, this did not take too much time and all evaluations were performed in a standardized way by rigorously trained evaluators. Subsequently, we only included healthy children. It might be interesting to evaluate a moderate-risk sample of e.g. infants born pre-term or low birth weight infants to obtain more information about the sensitivity of the Bayley-III-NL. The constitution of our sample also differs from the general constitution of the Flemish population. The results of our study might be biased for example by the high education level of the parents. Finally, despite the large sample size, there are still limited numbers of children within each age group. Therefore, we opted to make wider age groups of 3 months and not to use the original age groups of the Bayley-III-NL. Future research should focus on the evaluation of a large sample of typically developing children and children at moderate-risk with the Bayley-III-NL and the AIMS. As the age-range of the Bayley-III-NL is 16 days till 42 months, other evaluation tools should be used to assess validity in children older than 18 months. Also, the relation of the other Flemish Bayley-III-NL subscales, such as fine motor function, cognition and communication, with established developmental tools should be investigated.

5. Conclusions

The high correlation of the Bayley-III-NL GMS with the AIMS provides support for the concurrent validity of the Bayley-III-NL GMS, measuring gross motor development in infants. However, the Bayley-III-NL is better in differentiating the level of motor development than the AIMS in children with low and high scores. Our results on the divergent validity support that the Bayley-III-NL is able to measure distinct developmental domains separately and specifically. Secondly, population-specific reference values should be provided and used to avoid over- and underestimation of infant's development.

What this paper adds?

The Bayley-III-NL is a developmental evaluation tool for children up to 42 months. The AIMS is designed to evaluate motor development from 0 to 18 months. This paper provides support for the concurrent validity of the Bayley-III-NL Gross Motor Scale (GMS) and the AIMS. Given the high correlations, the AIMS can be used as an alternative test within routine neurological examination as it is based on observations, has few items, and requires less time to complete. The AIMS is designed from birth to independent walking and has shown to be most sensitive during the first 12 months. After the first year of life, a ceiling effect is seen which has a negative influence on the discriminative ability of the AIMS. At this moment, no ceiling effect has been reported yet for the Bayley-III-NL. Our results also show that the Bayley-III-NL GMS differentiates better the level of motor development in children with low and high scores. Secondly, low to moderate correlations were found between the other Bayley-III-NL subscales mutually (Fine Motor, Cognition, Receptive and Expressive Communication) and the AIMS, which confirms that these subscales measure distinct developmental

domains. Finally, this paper emphasizes the importance of population-specific reference values to avoid over- and under estimation of infant's development. There is still a need for new reference values in Flanders for the AIMS.

Declaration of interest

During data collection, training and testing materials were supported by Pearson BENELUX BV.

Acknowledgments

The authors thank the managers of the childcare centers, the home care providers and especially the parents of the children for their participation in this study. They also thank the Bayley-III-NL trainer, Selma Ruiter, who made the two-day education very instructive and interesting for the different evaluators. A special thank goes to Lori Cuveele and Linde Vandendriessche, two master students, Rehabilitation Sciences and Physiotherapy.

Appendix A. Description of selected items from Bayley-III-NL GMS in supine, prone, sitting and standing

	Item	Description
Supine	1	Thrusts legs in play
	2	Thrusts arms in play
	10	Holds head in midline
	20	Rolls from back to both sides
	24	Grasps foot with hands
	25	Rolls from back to stomach
Prone	5	Turns head to sides
	6	Makes crawling movements
	12	Controls head while prone series: 45°
	15	Elevates trunk while prone: Elbows and forearms
	17	Controls head while prone series: 90°
	18	Elevates trunk while prone series: Shifts weight
	21	Elevates trunk while prone series: Extended arms
	30	Crawl series: Crawls on stomach
	31	Crawls series: Crawl position
	32	Moves from sitting to hands and knees
Sitting	34	Crawls series: Crawl movement
	16	Sits with support: Short
	22	Sits without support: 5 s
	23	Pull to sit
	26	Sits without support: 30 s
	27	Sits without support and holds object
Standing	28	Rotates trunk while seated
	29	Makes walking movements
	33	Carries own weight
	35	Raises self to standing position
	36	Bends knees while standing
	37	Walks series: Walks with help
	38	Walks sideways with help
	39	Sits down with control
	40	Stands alone
	41	Stands up alone
	42	Walks series: Alone
	43	Walks series: Alone with coordination
	45	Squats without support
	46	Stands up: Advanced

References

- [1] N. Bayley, *Bayley Scales of Infant and Toddler Development-Third Edition*, TX: Harcourt Assessments, Inc., San Antonio, 2006.
- [2] M. Piper, J. Darrah, *Motor Assessment of the Developing Infant*, WB Saunders, Philadelphia, 1994.
- [3] A.L. Van Baar, L.J.P. Steenis, M. Verhoeven, D.J. Hessen, *Bayley-III-NL | Bayley Scales of Infant and Toddler Development - Third Edition - NL*, Pearson Assessment and Information, Amsterdam, The Netherlands, 2014.
- [4] P.L. de Albuquerque, M.Q. de F. Guerra, M. de C. Lima, S.H. Eickmann, Concurrent validity of the Alberta Infant Motor Scale to detect delayed gross motor development in preterm infants: a comparative study with the Bayley III, *Early Hum. Dev.* 10 (May) (2017) 1–13, <https://doi.org/10.1080/17518423.2017.1323974>.
- [5] E. Morales-Monforte, C. Bagur-calafat, N. Suc-lerin, E. Cazorla-sánchez, E. Cazorla-sa, The Spanish version of the Alberta Infant Motor Scale: validity and reliability analysis, *Dev. Neurorehabil.* 8423 (November) (2015) 1–7, <https://doi.org/10.3109/17518423.2015.1066461> (Early Online).
- [6] B.H. Connolly, N.O. McClune, R. Gatlin, Concurrent validity of the Bayley-III and the Peabody Developmental Motor Scale-2, *Pediatr. Phys. Ther.* 24 (4) (2012) 345–352, <https://doi.org/10.1097/PEP.0b013e318267c5cf>.
- [7] B.V. Acton, W.S.G. Biggs, D.E. Creighton, K. a H. Penner, H.N. Switzer, J.H.P. Thomas, ... Robertson, C.M. T, Overestimating neurodevelopment using the Bayley-III after early complex cardiac surgery, *Pediatrics* 128 (4) (2011) e794–e800, <https://doi.org/10.1542/peds.2011-0331>.
- [8] P.J. Anderson, C.R. De Luca, E. Hutchinson, G. Roberts, L.W. Doyle, Underestimation of developmental delay by the new Bayley-III Scale, *Arch. Pediatr. Adolesc. Med.* 164 (4) (2010) 352–356, [https://doi.org/10.1016/S8756-5005\(10\)79282-9](https://doi.org/10.1016/S8756-5005(10)79282-9).
- [9] T. Moore, S. Johnson, S. Haider, E. Hennessy, N. Marlow, Relationship between test scores using the second and third editions of the bayley scales in extremely preterm children, *J. Pediatr.* 160 (4) (2012) 553–558, <https://doi.org/10.1016/j.jpeds.2011.09.047>.
- [10] G. Reuner, A.C. Fields, A. Wittke, M. Löffprich, J. Pietz, Comparison of the developmental tests Bayley-III and Bayley-II in 7-month-old infants born preterm, *Eur. J. Pediatr.* 172 (3) (2013) 393–400, <https://doi.org/10.1007/s00431-012-1902-6>.
- [11] A.J. Spittle, P.J. Anderson, K.J. Lee, C. Ferretti, A. Eeles, J. Orton, ... Doyle, L. W, Preventive care at home for very preterm infants improves infant and caregiver outcomes at 2 years, *Pediatrics* 126 (1) (2010) e171–e178, <https://doi.org/10.1542/peds.2009-3137>.
- [12] Y.T. Yu, W.S. Hsieh, C.H. Hsu, L.C. Chen, W.T. Lee, N.C. Chiu, ... Jeng, S. F, A psychometric study of the Bayley Scales of Infant and Toddler Development - 3rd Edition for term and preterm Taiwanese infants, *Res. Dev. Disabil.* 34 (11) (2013) 3875–3883, <https://doi.org/10.1016/j.ridd.2013.07.006>.
- [13] L.J.P. Steenis, M. Verhoeven, D.J. Hessen, A.L. Van Baar, Performance of Dutch children on the Bayley III: a comparison study of US and Dutch norms, *PLoS One* 10 (8) (2015) 1–13, <https://doi.org/10.1371/journal.pone.0132871>.
- [14] E.A. Cromwell, Q. Dube, S.R. Cole, C. Chirambo, A.E. Dow, R.S. Heyderman, A. Van Rie, Validity of US norms for the Bayley Scales of Infant Development-III in Malawian children, *Eur. J. Paediatr. Neurol.* 18 (2) (2014) 223–230, <https://doi.org/10.1016/j.ejpn.2013.11.011>.
- [15] A. De Kegel, W. Peersman, K. Onderbeke, T. Baetens, I. Dhooge, H. Van Waelvelde, New reference values must be established for the Alberta Infant Motor Scales for accurate identification of infants at risk for motor developmental delay in Flanders, *Child Care Health Dev.* 39 (2) (2013) 260–267, <https://doi.org/10.1111/j.1365-2214.2012.01384.x>.
- [16] K.M.W. Fleuren, L.S. Smit, T. Stijnen, A. Hartman, New reference values for the Alberta Infant Motor Scale need to be established, *Acta Paediatr.* 96 (3) (2007) 424–427, <https://doi.org/10.1111/j.1651-2227.2007.00111.x>.
- [17] D.E. Hinkle, W. Wiersma, *Applied Statistics for Behavioral Sciences*, International 5th Edition, Cengage Learning, Inc., Boston, 2009.
- [18] K.M. Almeida, M.V.P. Dutra, R.R. De Mello, A.B.R. Reis, P.S. Martins, Concurrent validity and reliability of the Alberta Infant Motor Scale in premature infants, *J. Pediatr.* 84 (5) (2008) 442–448, <https://doi.org/10.2223/JPED.1836>.
- [19] S.F. Jeng, K.I. Yau, L. Chen, S.F. Hsiao, Alberta Infant Motor Scale: reliability and validity when used on preterm infants in Taiwan, *Phys. Ther.* 80 (2) (2000) 168–178.
- [20] Z.H. Yıldırım, N. Aydınli, B. Ekici, B. Tath, M. Çalışkan, Can Alberta infant motor scale and milani comparetti motor development screening test be rapid alternatives to bayley scales of infant development-II at high-risk infants, *J. Indian Acad. Neurol.* 15 (3) (2012) 196–199.
- [21] P.L. De Albuquerque, A. Lemos, M.Q.D.F. Guerra, S.H. Eickmann, Accuracy of the Alberta Infant Motor Scale (AIMS) to detect developmental delay of gross motor skills in preterm infants: a systematic review, *Dev. Neurorehabil.* 8423 (October) (2014) 1–7, <https://doi.org/10.3109/17518423.2014.955213>.
- [22] D. Campos, V.M.G. Gonçalves, M.M. Guerreiro, D.C. Santos, M.M.F. Goto, A.V. Arias, T.M. Campos-Zanelli, Comparison of motor and cognitive performance in infants during the first year of life, *Pediatr. Phys. Ther.* 24 (2) (2012) 193–197, <https://doi.org/10.1097/PEP.0b013e31824d2db7> (discussion 198).
- [23] OECD Family Database: Enrolment in Childcare and Pre-school, (2016), p. 8 Retrieved from https://www.oecd.org/els/soc/PF3_2_Enrolment_childcare_preschool.pdf.
- [24] M.P.C.M. Luijk, M. Linting, J. Henrichs, M. Herba, M.L. Verhage, J.J. Schenk, ... van Ijzendoorn, M. H, Hours in non-parental child care are related to language development in a longitudinal cohort study, *Child Care Health Dev.* 41 (6) (2015) 1188–1198, <https://doi.org/10.1111/cch.12238>.
- [25] R. Lekhal, H.D. Zachrisson, M.V. Wang, S. Schjolberg, T. Von Soest, Does universally accessible child care protect children from late talking? Results from a Norwegian population-based prospective study, *Early Child Dev. Care* 181 (2011) 1007–1019.
- [26] NICHD ECCRN, Type of child care and children's development at 54 months, *Early Child Res. Q.* 19 (2004) 203–230.
- [27] J. Bradshaw, C. Skinner, W. Van Lancker, Coverage of childcare in the European Union, *ESPN Flash Report*, 66, 2015.
- [28] J. Darrah, D. Bartlett, T.O. Maguire, W.R. Avison, T. Lacaze-Masmonteil, Have infant gross motor abilities changed in 20 years? A re-evaluation of the Alberta Infant Motor Scale normative values, *Dev. Med. Child Neurol.* 56 (9) (2014) 877–881, <https://doi.org/10.1111/dmcn.12452>.
- [29] A. Koren, S.M. Reece, L. Kahn-D'angelo, D. Medeiros, Parental information and behaviors and provider practices related to tummy time and back to sleep, *J. Pediatr. Health Care* 24 (4) (2010) 222–230, <https://doi.org/10.1016/j.pedhc.2009.05.002>.
- [30] A. Majnemer, R.G. Barr, Association between sleep position and early motor development, *J. Pediatr.* 149 (5) (2006) 623–629, <https://doi.org/10.1016/j.jpeds.2006.05.009>.
- [31] B.E. Davis, R.Y. Moon, H.C. Sachs, M.C. Ottolini, Effects of sleep position on infant motor development, *Pediatrics* 102 (5) (1998) 1135–1140, <https://doi.org/10.1542/peds.102.5.1135>.
- [32] J.W. Jantz, C.D. Blosser, L.A. Fruechting, A motor milestone change noted with a change in sleep, *Arch. Pediatr. Adolesc. Med.* 151 (6) (1997) 565–568.
- [33] J.S. Salls, L.N. Silverman, C.M. Gatty, The relationship of infant sleep and play positioning to motor milestone achievement, *Am. J. Occup. Ther.* 56 (5) (2002) 577–580, <https://doi.org/10.5014/ajot.56.5.577>.
- [34] S. Chinta, K. Walker, R. Halliday, A. Loughran-Fowlds, N. Badawi, A comparison of the performance of healthy Australian 3-year-olds with the standardised norms of the Bayley Scales of Infant and Toddler Development (version-III), *Arch. Dis. Child.* (2014) 621–624, <https://doi.org/10.1136/archdischild-2013-304834> (Bayley III).