



# Parental and professional assessment of early child development: The ASQ-3 and the Bayley-III-NL

Leonie J.P. Steenis<sup>a,\*</sup>, Marjolein Verhoeven<sup>a</sup>, Dave J. Hessen<sup>b</sup>, Anneloes L. van Baar<sup>a</sup>

<sup>a</sup> Child and Adolescent Studies, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands

<sup>b</sup> Methods and statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands

## ARTICLE INFO

### Article history:

Received 22 August 2014

Received in revised form 17 December 2014

Accepted 17 January 2015

### Keywords:

Infant and toddler development

ASQ-3

Bayley-III-NL

Sensitivity

Specificity

Screening

## ABSTRACT

**Background:** The Ages and Stages Questionnaire (ASQ-3) is frequently used for screening developmental delay in problem solving, communication, fine- and gross motor skills and personal-social behavior of infants, toddlers and preschool aged children. The adequacy of the ASQ-3 is evaluated for Dutch children by comparing results of the ASQ-3, completed by parents, to results of a standardized, professionally administered developmental assessment of cognition, fine- and gross motor skills and receptive and expressive communication for infants and toddlers: the Bayley-III-NL.

**Methods:** The ASQ-3 and Bayley-III-NL were administered to 1244 children aged 1 to 43 months old. Two age cohorts were used: 1) the 2–16 month age-versions; and 2) the 18–42 month age-versions. Cutoff points for all ASQ-3 age-versions were calculated in three ways. Sensitivity and specificity of the ASQ-3 were evaluated with four methods, using different cutoff point combinations of 1 SD or 2 SD below the mean.

**Results:** Overall, sensitivity was between 7% and 77% and specificity between 53% and 99%. Sensitivity and specificity values were higher for the older age-cohort than for the younger age-cohort. For the older age-cohort, the best sensitivity (69%) and specificity (92%) was found, using 1 SD for the total ASQ-3 score and 2 SD for the Bayley-III-NL subtests as cutoff points.

**Conclusions:** For the oldest age-cohort, the ASQ-3 for now has the best potential as a screener for Dutch children. The ASQ-3 identifies most children without a developmental delay according to the Bayley-III-NL, but sensitivity needs improvement.

© 2015 Elsevier Ireland Ltd. All rights reserved.

## 1. Introduction

Children experiencing a developmental delay benefit from early intervention [1,2] and therefore it is crucial to detect a developmental delay as early as possible. Despite professional monitoring systems focusing on detection of developmental delay in early childhood, only 30% of the children with developmental problems in the United States are identified before they reach school age [3]. As previous studies indicate that parents can be reliable and valid sources of information on the development of their own child, current monitoring systems might benefit from including knowledge of parents about their child's development [4–8]. A worldwide, frequently used parental screening tool with good psychometric characteristics is the Ages and Stages

Questionnaire (ASQ-3; [9]). The ASQ-3 identifies infants and toddlers between one month and 66 months of age, who are at risk of a developmental delay regarding problem solving, communication, fine motor skills, gross motor skills and personal social behavior and is developed and validated in the US. The ASQ-3 consists of 21 age-versions and each age-version consists of six questions per developmental domain concerning age-specific developmental key-milestones.

All age-versions of the ASQ-3 have been translated into Dutch for potential use as a screening measure in the Netherlands as well. Recently, the Bayley III, an extensive professional assessment of developmental level in cognition, communication and motor functioning of children aged between two weeks and 42 months and 15 days old, has been translated, normed and validated for Dutch children: The Bayley-III-NL [10]. By comparing the answers of the parents on the Dutch ASQ-3 with the results of the professionally administered Bayley-III-NL in a Dutch sample, the adequacy of the screening tool for Dutch children can be studied.

A good screening instrument should be able to identify at least 70% of the children that show a developmental delay on an extensive assessment (sensitivity), and it should also identify at least 70% of the children who perform within the normal range according to an extensive developmental assessment (specificity) [11]. The results

**Abbreviations:** ASQ, Ages and Stages Questionnaire; ASQ-2, Ages and Stages Questionnaire—second edition; ASQ-3, Ages and Stages Questionnaire—third edition; Bayley III, Bayley scales of infant and toddler development—third edition; Bayley-III-NL, the Dutch version of the Bayley scales of infant and toddler development—third edition; BSID-II, Bayley Scales of Infant Development—second edition; NPV, negative predictive value; PPV, positive predictive value; SD, standard deviation(s).

\* Corresponding author. Tel.: +31 302531196.

E-mail addresses: [l.j.p.steenis@uu.nl](mailto:l.j.p.steenis@uu.nl) (L.J.P. Steenis), [j.c.t.verhoeven@uu.nl](mailto:j.c.t.verhoeven@uu.nl) (M. Verhoeven), [d.j.hessen@uu.nl](mailto:d.j.hessen@uu.nl) (D.J. Hessen), [a.l.vanbaar@uu.nl](mailto:a.l.vanbaar@uu.nl) (A.L. van Baar).

of earlier studies comparing the ASQ (second and third version [9, 12]) with the Bayley scales to assess developmental level of infants and toddlers (BSID II and Bayley III; [13,14]), showed that the sensitivity and specificity values of the ASQ vary substantially across studies (see Table 1). An explanation for this variation is that these studies differed in design: Different cutoff points to determine developmental delay were used; different combinations of cutoff points per instrument were used; the samples of the studies differed regarding high or low risk samples; different versions of the Bayley scales were used; and the age-versions of the ASQ that were evaluated, differed per study.

Another explanation for the different results of the studies described in Table 1 is that most studies were conducted in other countries than the US, but still used the original ASQ-3 as it is developed and validated in the US. However, earlier studies showed the differences between the development of children in the US and children in other Western countries [15–19]. Because the constellation of the items and the cutoff points per age-version of the ASQ-3 are based on the development of children in the US population, the ASQ-3 might not be adequate for use in other countries. As a result, sensitivity and specificity values calculated in a US sample, might differ from sensitivity and specificity values calculated in non-US samples.

Currently, it is unknown whether the translated ASQ-3 is adequate for use as a screening measure in the Netherlands for infants and toddlers. The aim of this study is to evaluate in a community sample whether the Dutch translation of the ASQ-3 is an accurate screening measure for infants and toddlers at risk of a delay in the Netherlands for the age-versions of 2 months to 42 months, by comparing its results with the results on the Bayley-III-NL. As it can be expected that in a community sample only small numbers of children will show a developmental delay in a specific age group, two age cohorts are evaluated: one younger and one older than 18 months.

## 2. Methods

### 2.1. Participants

Parents of children between the ages of 2 weeks to 42 months and 15 days were asked to participate in a study on early child development aiming at a representative sample for the Dutch population (see Table 2 for stratification target percentages in the Dutch population). Both healthy children and children with known risk factors for having a developmental delay or who were already diagnosed with a developmental delay (e.g. preterm born children and children small for gestational age) were included in the study. In total, Bayley-III-NL data were available for 2053 children and 1912 (96.5%) of these children had complete and reliable Bayley-III-NL results. For 1378 of these children, ASQ-3 questionnaires were available, of which 1244 (90.3%) were completed at the correct age of the child given the age-ranges of the ASQ-3 age-versions. Children were between 16 days and 44 months old (mean = 16.93 months, SD = 12.20) at time of completion of the ASQ-3. Of these children, 7.4% had a known risk or diagnosis regarding developmental delay. Table 2 presents the background characteristics for the two age-cohorts separately.

### 2.2. Instruments

#### 2.2.1. The ASQ-3

The ASQ-3 is a screening questionnaire completed by parents, which identifies children in the age range of 1 month to 60 months, at risk of a developmental delay [9]. It consists of 21 age versions, of which the 18 versions between 2 months and 42 months were used in this study. The ASQ-3 concerns five developmental domains: Communication, Gross Motor, Fine Motor, Problem Solving and Personal Social. Next to that, open questions about the general health of the child are included and parents can indicate whether they have concerns on the development of their child. Each domain contains six questions about important

developmental milestones appropriate for the age-version at hand. The response of the parents to each item is indicated as “yes” (score = 10), “sometimes” (score = 5), or “not yet” (score = 0). When one or two items per domain are missing, these missing values can be replaced by the average score of that child on that domain. Raw scores are calculated for each developmental domain by summing the scores of the six questions. The total score of each domain ranges from 0 to 60. The manual advises that a child with scores between 1 SD and 2 SD below the normative mean on any domain, be provided with developmental activities, after which the child should be rescreened or referred for further evaluation. When a child scores 1 SD below the normative mean on two or more domains, or 2 SD below the normative mean on at least one domain, immediate referral for further assessment is advised. Psychometric characteristics of the ASQ-3 were found to be excellent [12]. The ASQ-3 was translated into Dutch and checked by back translation for the 18 age-versions suitable for 2 to 42 month old children. In addition, experts in child development evaluated the translation of the items.

#### 2.3. The Bayley-III-NL

The Bayley scales of infant and toddler development—third version (Bayley III) form a professionally administered instrument that assesses the developmental level of children between 16 days and 42 months plus 15 days old [14]. For administration purposes, this age-range is divided into 17 age groups. The Bayley III consists of five scales: Cognition Scale (91 items), Fine Motor Subtest (66 items), Gross Motor Subtest (72 items), Receptive Communication Subtest (49 items), and Expressive Communication Subtest (48 items). The Bayley III has been translated into Dutch, adapted for Dutch culture and checked by back translation, in order to develop the Bayley-III-NL. Experts reviewed the translation and adaptation of the items [10].

Recently, Dutch norms were created and just as in the original US version, the subtests of the Bayley-III-NL have standardized scores with a range of 1 to 19 and a mean of 10 and a standard deviation of 3. Scaled scores below 7 correspond to 1 SD below the mean; scaled scores below 4 correspond to 2 SD below the mean. It is advised to refer a child for treatment when it scores 1 SD below the normative mean on two or more subtests or when it scores 2 SD below the mean on at least one subtest. The Bayley-III-NL was found to be a reliable and valid instrument with good psychometric characteristics for Dutch children. The reliabilities varied for the five subtests from .82 to .92.

#### 2.4. Procedure

Children were recruited via day-care centers, advertisements in newspapers, through personal connections and via child health centers in the Netherlands. Parents were asked questions about the general health, birth weight and gestational age of the child to determine whether the child belonged to the group of children at risk of a developmental delay. This group was allowed to be a maximum of 10% of the total sample [10].

Two weeks prior to an appointment for the assessment with the Bayley-III-NL, parents received the age-appropriate questionnaire of the ASQ-3 and an informed consent form by mail, which they were asked to complete at home. During the visit, the ASQ-3 and the signed informed consent form were collected and the mother was asked if she experienced any difficulties completing the questionnaire. Mothers that forgot to complete the questionnaire were once again asked to complete the questionnaire at home and send it by mail to the research team. In total, 20.2% of the ASQ-3 questionnaires were completed at the same day as the Bayley-III-NL assessment, 68.5% were completed before the day of test administration with a mean of 7.3 days (range of 1 to 70 days with 90% completion within 6–8 days) and 11.2% after the day of test administration with a mean of 8.6 days (range of 1 to 65 days with 90% completion between 8 and 12 days).

**Table 1**  
Studies comparing the ASQ and the Bayley scales.

Authors	Sample	Bayley scales version	ASQ version	Norms	Criterion for delay		Cut-off points used		Results			
					BSID-II (Number of scales)	ASQ (Number of domains)	BSID-II ( $\leq ?SD$ )	ASQ ( $\leq ?SD$ )	ASQ age-version	N	Sens. (%)	Spec. (%)
Bian et al., 2012 [20]	Low risk Chinese sample	II	3	BSID-II: US ASQ: Chinese	$\geq 1$	$\geq 1$	1.5	2	6	52	83	93
									12	60	71	81
									18	52	100	84
									24	49	80	84
Gollenberg et al., 2010 [21]	Low risk US sample	II	2	US	$\geq 1$	$\geq 1$	1	1	30	42	100	85
									24	53	39	87
Simard et al., 2012 [22]	Canadian sample Gestational age: 29–36 6/7 weeks Birth weight: <2500 g	II	2	US	$\geq 1$	$\geq 1^*$	2	2	24	53	100	93
									24	112	92	55
							1 MDI	1	12	124	60	68
									24	112	92	55
							1 PDI	1	12	124	52	90
									24	112	50	73
							1 MDI	1.5	12	124	45	78
									24	112	88	72
							1 PDI	1.5	12	124	39	96
									24	112	50	73
Schonhaut et al., 2013 [23]	Chilean combined term, late preterm and extreme preterm sample	III	3	US	$\geq 1$	$\geq 1$	1	2	12	124	20	88
									24	112	75	78
							1 PDI	2	12	124	25	97
									24	112	31	92
Skellern et al., 2001 [24]	Australian sample Ex-premature infants <31 weeks gestation	II	1	US	Delay on MDI	$\geq 1$	1	2	8	110	71	76
									18	100	79	84
									30	96	82	84
Smith et al., 2012 [25]	Australian infants After surgery without surgery, and control group	III	2	US	Delay on Fine Motor Subtest	Fine Motor domain	1	1	18	24	50	91
									12	332	20	98
Woodward et al., 2011 [26]	US sample Birth weight 500–999 g Mechanical ventilation 12 h–24 h after birth	II	2	US	$\geq 1$	$\geq 1$	1	2	18–22	219	63	75
									2		95	32
									2		73	65

Note. Sens. = sensitivity; Spec. = specificity; MDI = Mental Developmental Index; PDI = Psychomotor Developmental Index.

\* Within self-made categories.

**Table 2**  
Background information.

	ASQ-3 version 2–16	ASQ-3 version 18–42	Dutch population **
N	679	565	
Mean age (SD)	7.25 (4.6)	28.56 (7.4)	–
Boys (%)	53.3	49.2	51.0
Gestational age in weeks, mean (SD)	39.5 (1.6)	39.49 (2.0)	–
Range	28–43	27–43	–
Birth weight in grams, mean (SD)	3496 (558.3)	3438 (587.6)	–
Range	1070–5020	745–4950	–
Ethnicity mother: Dutch (%)	79.2	84.1	75.0
Educational level mother*			
Low (%)	10.9	16.0	11.5
Middle (%)	37.7	39.0	38.2
High (%)	51.7	45.0	50.3

\* 'Low educational level' refers to special education, primary school, or pre-vocational secondary education (<12 years); 'middle educational level' refers to senior general secondary education, pre-university education, or secondary vocational education (13–16 years); 'high educational level' refers to higher professional education or university (17+ years).

\*\* Percentages retrieved from the Central Bureau of Statistics [32].

The Bayley-III-NL was administered at locations within acceptable traveling distance for the child and free from distracting stimuli. Because young infants are only awake and alert for small periods of time and traveling to a lab could be too fatiguing, children up to six months of age were tested at home. Examiners were experienced clinicians or students in the final year of their bachelor- or master education, and all were trained to be reliable in their test administration. All examiners also scored a Bayley-III-NL assessment on film and had to acquire an inter-rater reliability level with a minimal consensus rate of at least 80% per subtest. Their scores were compared to the scores of the trainer and the average kappa for all administered items over all subtests was .77 (SD = .05). Consecutive teams of examiners administered the Bayley-III-NL. As such, inter-rater reliability was established throughout the study. Examiners were blind to the ASQ-3 results.

#### 2.4.1. Data analysis

All analyses were performed using SPSS 20. First, we examined the presence of missing data for the ASQ-3. When one or two items were missing per domain, the guidelines as provided in the ASQ-3 manual were followed by replacing missing values by the average score of that child on that particular developmental domain. When 3–6 items per domain were missing, missing data were analyzed with Little's MCAR test in relation to the background variables age, gender, region, child's health, ethnicity of the parents per developmental domain and per age-version of the ASQ-3. When the Little's MCAR tests per domain showed that data was missing completely at random, missing values were imputed by means of the Expectation Maximization method in SPSS [27].

Second, for each child we determined whether it had a developmental delay (based on Bayley-III-NL scores) or a risk at a developmental delay (based on ASQ-3 scores). For the Bayley-III-NL, the scaled scores based on the Dutch norms were used [10]. For the ASQ-3, however, there are currently no Dutch norms available. As earlier studies indicated that US norms might not be adequate to use in other countries [15–19], we calculated cutoff scores based on the data of the current sample using three approaches. For each approach, the cutoff points were calculated per age-version, given that each age version of the ASQ-3 consists of a unique set of age-appropriate questions and so the cutoff points should also be specific for each age-version. The first two approaches were based on the conventional method of using 1 SD or 2 SD as cutoff point. This approach is also used in most studies described in Table 1. For the first approach, raw scores of each developmental domain were standardized separately for each age-version of the ASQ-3. The z-scores

showed which children scored 1 SD or 2 SD below the mean in relation to the score of their age-group. For the second approach, total ASQ-3 scores were calculated by summing the total score of 30 items (6 items × 5 domains, total scores ranging between 0 and 300) and these total scores were standardized separately for each age-version of the ASQ-3.

In the third approach, Receiver Operating Characteristic Curve (ROC curve) analyses were performed in order to determine the most optimal cutoff points with the best sensitivity and specificity values. In order to perform ROC curve analyses, a dichotomized score based on a "gold standard"—in our study the Bayley-III-NL—is necessary to use as a reference. As the personal-social domain has no conceptual match with the Bayley-III-NL subtests, ROC curve analyses could only be performed for the total ASQ-3 score and not for the ASQ-3 domains separately. Based on their Bayley-III-NL scaled scores, children were categorized as being developmentally delayed when they scored at least 1 SD or 2 SD below the mean on at least one subtest as cutoff point. An area under the curve (AUC) of .7 or higher reflects a fair to good accuracy in the screening capacities of the ASQ-3. An AUC smaller than .7 indicates poor accuracy in the screening capacities of the ASQ-3 and that no valuable cutoff point can be determined [28].

Finally, sensitivity, specificity, negative predictive values (NPV) and positive predictive values (PPV) for the ASQ-3 were calculated using four different combinations of the cutoff points described above. Sensitivity and specificity values above 70% are considered sufficient [11]. In view of the small number of children that are expected to show a developmental delay in a specific age group, all sensitivity, specificity, NPV and PPV were calculated for two age cohorts: 1) a young age-cohort with infants that received the 2–16 month ASQ-3 version and; 2) an older age-cohort with toddlers that received the 18–42 month ASQ-3 version. We choose to split the sample at the 18 month ASQ-3 version, as this is a critical age for determining certain developmental delays. For instance, a child is considered a "late talker" when it does not speak yet at the age of 18 months [29] and is considered a late walker when it does not walk independently at the age of 18 months [30].

In earlier studies and the manuals of the ASQ-3 and the Bayley-III-NL, four comparison methods were found for the evaluation of the sensitivity and specificity of the ASQ-3. In the most frequently used comparison method (see Table 1), the cutoff points of both the ASQ-3 and Bayley-III-NL were respectively 1 SD or 2 SD below the mean on at least one subdomain, resulting in the following comparisons: 1 SD–1 SD; 1 SD–2 SD; 2 SD–1 SD; 2 SD–2 SD (comparison method 1).

The second comparison method is based on the referral guidelines of the ASQ-3 and the Bayley-III-NL: A child should be referred when it has a score which is at least 1 SD below the mean on two or more ASQ-3 domains or Bayley-III-NL subtests, or a score that is lower than 2 SD below the mean on at least one domain or subtest. Thus, whereas in the first comparison method children scoring 1 SD and children scoring 2 SD below the mean on at least one domain were examined separately, in this second method only one group of children is examined: i.e. children who either score 1 SD below the mean on at least 2 domains or who score 2 SD on at least 1 domain. (Comparison method 2.)

The third and fourth comparison methods used cutoff scores based on the total score for the ASQ-3. In the third comparison method, sensitivity, specificity, NPV, and PPV were calculated using 1 SD and 2 SD as cutoff point for the total ASQ-3 score and at least 1 SD or 2 SD below the mean on at least one Bayley-III-NL subtest. The results of the scores on the ASQ-3 and the Bayley-III-NL were compared for different combinations of cutoff points respectively: 1 SD–1 SD; 1 SD–2 SD; 2 SD–1 SD; 2 SD–2 SD (comparison method 3). In the fourth comparison method, ROC curve cutoff points were used for the total ASQ-3 score and 1 SD or 2 SD below the mean on at least one subtest was used for the Bayley-III-NL. This comparison is only useful when AUC was .7 or higher for all age-versions in a particular age-cohort.



### 3. Results

#### 3.1. Preliminary results

Parents might have only participated in this study because they are concerned regarding their child's development, leading to a selection bias of the sample. This does not seem the case, as only 168 (13.5%) out of the 1244 participating parents indicated such concerns in the open ASQ-3 questions. Also the fact that not for all children the ASQ-3 and Bayley-III-NL were completed in the same order might have led to bias in the final results of our study. However, MANOVA analyses conducted per age-version showed that the order in which the tests were completed (ASQ-3 prior to Bayley-III-NL or vice versa) did not lead to differences in test scores.

After replacing the missing values when one or two items were missing per domain, 0.2–1.8% of the data was missing per ASQ-3 item. Little's MCAR test revealed that per ASQ-3 domain, data was missing completely at random in relation to the background variables. Evaluating the missing data per ASQ-3 age-version, Little's MCAR test revealed that 82% of the missing data was missing completely at random. Of the 43 children (3% of 1244 children) that had missing data, 33% had a score below 1 SD on at least one subtest of the Bayley-III-NL. All missing data were imputed using the Expectation Maximization method in SPSS. To study whether this imputation influenced the results, all analyses were also performed without the imputed ASQ-3 values. As the results remained the same, we only present the results based on the imputed data.

#### 3.2. Cutoff points

Table 3 presents the means, standard deviations and cutoff points per ASQ-3 domain and the total ASQ-3 score. As can be seen in this table, the means and SD's of the ASQ-3 differ largely across developmental domains and age versions. Moreover, for the Gross Motor domain, the cutoff score of 2 SD is negative for three age versions (8, 9 and 12), making it impossible for children to score 2 SD or lower on this developmental domain. Table 4 presents the means and standard deviation per ASQ-3 age-version of the Bayley-III-NL subtests.

##### 3.2.1. Comparison of ASQ-3 and Bayley-III-NL results

Table 5 presents the sensitivity, specificity, NPV and PPV for three comparison methods of the ASQ-3 and the Bayley-III-NL for both age cohorts with ASQ-3 age-versions 2–16 months and 18–42 months.

##### 3.2.2. Comparison method 1: indications for developmental delay on at least one ASQ-3 domain and on at least one Bayley-III-NL subtest using Z-scores

For the youngest age-cohort, the sensitivity was low to moderate ranging from 19.2% to 57.9%. Using 2 SD below the mean as cutoff point for the ASQ-3 and 1 or 2 SD below the mean for the Bayley-III-NL, resulted in high specificity values of respectively 89.8% and 86.7% for the age-cohort of 2–16 months old. For the oldest age-cohort, one analysis showed a sufficient sensitivity of 76.9%, for the combination of a score below 1 SD on the ASQ-3 and below 2 SD on the Bayley-III-NL. This analysis however, resulted in a low specificity of 62.7%. For the other SD cutoff combinations in the age-cohort from 18–43 months, sufficient specificity values were found in combination with low to moderate sensitivity values ranging from 27.3% to 62.0%.

##### 3.2.3. Comparison method 2: indications of developmental delay based on the referral guidelines using Z-scores

For both age-cohorts, specificity values were above 70% and therefore sufficient, but a low sensitivity value of 38.0% was found for the youngest age-cohort and a moderate sensitivity of 62.7 was found for the oldest age-cohort.

##### 3.2.4. Comparison method 3: indications for developmental delay based on the total score for the ASQ-3 using Z-scores

For both age cohorts, the sensitivity values were low to moderate ranging from 7.2% to 69.2% for all cutoff point combinations with sufficient specificity values. A combination of a cutoff point of 1 SD below the mean on the ASQ-3 and 2 SD below the mean on the Bayley-III-NL for the oldest age cohort, resulted in the best outcome of all three methods with a sensitivity of 69.2%, a specificity of 92.4%; negative (NPV) and positive predictive values (PPV) were 99.2% and 17.6%, respectively.

##### 3.2.5. Comparison method 4: indications for developmental delay based on the total score for the ASQ-3 using ROC curve cutoff points

When performing ROC-curve analyses to determine cutoff points for the total ASQ-3 score per age-version we came across two problems: 1) ROC-curves could not be analyzed for all age-versions of the ASQ-3, as not in every age group children scored below 1 SD or 2 SD on one or more subtests of the Bayley-III-NL; 2) the area under the curve (AUC) was below .7 for several age-groups (see Table 3). Therefore the comparison method using cutoff points determined by ROC curve analyses could not be performed.

#### 3.3. Overall results

Regarding sensitivity, all three comparison methods showed low to adequate values ranging from 7.3% to 76.9% and moderate to adequate specificity values ranging from 59.3% to 99.3%. Overall, the sensitivity and specificity values were better for the oldest age cohort with the 18–42 month ASQ-3 age versions. The best sensitivity-specificity combination was found for the combination of 1 SD as cutoff point for the ASQ-3 and 2 SD as cutoff point for the Bayley-III-NL for the 18–42 month ASQ-3 age versions.

### 4. Discussion

In this study, the adequacy of the ASQ-3 as a screening instrument for Dutch children between 2 weeks and 42 months old was examined. Scores of the ASQ-3 and the Bayley-III-NL were compared in a large sample of Dutch infants for two age cohorts; one cohort of infants for whom the 2–16 month ASQ-3 age versions were completed and another cohort of toddlers for whom the 18–42 month ASQ-3 age versions were completed. We used three different comparison methods, in which standardized scores of the ASQ-3 served as cutoff points: 1) indications for developmental delay on at least one ASQ-3 domain and one Bayley-III-NL subtest; 2) indications of developmental delay based on the referral guidelines for both measures; and 3) indications for developmental delay based on the total score for the ASQ-3 using Z-scores and on at least one subtest of the Bayley-III-NL. A fourth comparison method, in which cutoff points for the ASQ-3 were based on ROC-curve analyses could not be performed, as ROC cutoff points could not be determined for all age-versions of the ASQ-3. The three comparison methods showed low to moderate sensitivity values ranging from 7.2% to 76.9%. This indicates that many children who according to the ASQ-3 were at risk for a developmental delay, were not identified as such by the Bayley-III-NL. All specificity values were moderate to high ranging from 53.2% to 98.4%, indicating that most children were not identified as at risk for delay by the ASQ-3 indeed did not show a delay on the Bayley-III-NL. Overall, the sensitivity and specificity values were generally higher for the older age-cohort than for the younger age-cohort, and thus it seems that the ASQ-3 is better able to screen older children. Unfortunately, none of the comparison methods resulted in a sufficient percentage of 70% for sensitivity and specificity concurrently.

The best sensitivity-specificity combination of respectively 69.2% and 92.4% was found for the 18–42 cohort, using comparison method 3 with a cutoff point of 1 SD for the total ASQ-3 score and a cutoff

**Table 3**  
Cutoff points for the ASQ-3 per age-version determined using 3 different approaches.

		ASQ-3 age-versions																	
		2	4	6	8	9	10	12	14	16	18	20	22	24	27	30	33	36	42
Per ASQ-3 domain																			
Approach 1																			
Communication	Mean	41.14	49.19	45.00	48.53	35.99	38.18	38.79	38.49	36.13	30.65	41.78	41.16	50.24	50.10	56.31	54.66	53.84	52.32
	SD	15.87	7.51	9.56	9.72	13.51	10.86	13.20	12.53	13.13	13.23	14.79	15.07	12.88	12.49	7.60	7.79	5.27	7.89
	Cutoff 1 SD	25.27	41.86	35.44	38.81	22.48	27.32	25.59	25.96	23.00	17.42	26.99	26.09	37.36	37.61	48.71	46.87	48.57	44.43
	Cutoff 2 SD	9.40	34.17	25.88	29.09	8.97	16.46	12.39	13.43	9.87	4.19	12.20	11.02	24.48	25.12	41.11	39.08	43.3	36.54
Gross Motor	Mean	52.65	51.46	31.48	31.71	27.33	37.28	32.18	39.53	44.78	42.92	44.29	45.32	52.21	44.34	51.11	52.24	54.53	51.95
	SD	8.31	9.40	13.75	16.58	18.56	18.04	19.87	18.61	18.36	17.47	13.31	15.26	9.92	12.78	9.28	8.55	7.35	7.58
	Cutoff 1 SD	44.34	42.06	17.73	15.13	8.77	19.24	12.31	20.92	26.42	25.45	30.98	30.06	42.29	31.56	41.83	43.69	46.98	44.37
	Cutoff 2 SD	36.03	32.66	3.98	-1.45	-9.79	1.20	-7.56	2.31	8.06	7.98	17.67	14.80	132.37	18.78	32.55	35.14	39.63	36.79
Fine Motor	Mean	42.36	45.04	41.89	45.92	50.89	47.75	45.34	41.49	45.48	47.98	52.62	46.70	51.00	43.11	45.19	46.32	50.46	49.23
	SD	11.03	11.52	13.02	14.61	12.05	12.40	10.13	12.26	11.09	11.31	7.40	9.21	7.43	11.50	14.37	13.45	10.84	12.47
	Cutoff 1 SD	31.33	33.52	28.87	31.31	38.84	35.35	35.21	29.23	34.39	36.67	45.22	37.49	43.57	31.61	30.82	32.87	39.62	36.76
	Cutoff 2 SD	20.30	22.00	15.85	16.70	26.79	22.95	25.08	16.97	23.30	25.36	37.82	28.28	36.14	20.11	16.45	19.42	28.78	24.29
Problem Solving	Mean	41.33	50.26	47.29	47.94	44.64	45.35	41.85	43.15	46.87	41.06	42.57	46.96	46.04	48.98	49.96	53.26	54.36	52.91
	SD	12.21	9.86	12.85	11.19	10.70	11.68	13.50	13.01	12.17	12.40	8.02	10.96	9.11	9.85	10.03	8.79	8.38	10.05
	Cutoff 1 SD	29.12	40.40	34.44	36.75	33.94	33.67	28.35	30.14	34.70	28.66	34.55	36.00	36.93	39.13	39.93	44.47	45.98	42.86
	Cutoff 2 SD	16.91	30.54	21.59	25.56	23.24	21.99	14.85	17.13	22.53	16.26	26.53	25.04	27.82	29.28	29.90	35.68	37.60	32.81
Personal Social	Mean	44.46	47.38	39.37	47.19	43.31	40.55	41.13	37.60	39.16	40.42	46.33	49.19	45.87	43.65	50.92	52.23	53.29	50.93
	SD	10.60	10.65	13.15	9.88	11.81	11.92	10.67	13.15	12.14	13.44	11.37	8.57	9.06	11.23	7.45	9.18	5.19	7.79
	Cutoff 1 SD	33.86	36.73	26.22	37.31	31.50	28.63	30.46	24.45	27.02	26.98	34.96	40.62	36.81	32.42	43.47	43.05	48.10	43.14
	Cutoff 2 SD	23.26	26.08	13.07	27.43	16.69	16.71	19.79	11.30	14.88	13.54	23.59	32.05	27.75	21.19	36.02	33.87	42.91	35.35
ASQ-3 total score																			
Approach 2																			
z-score	Mean	221.95	243.34	205.02	221.33	202.16	209.10	199.29	200.27	212.54	203.02	227.58	229.34	245.35	230.18	253.47	258.70	266.38	257.35
	SD	41.55	35.84	45.92	42.49	46.33	45.05	45.46	47.72	46.82	50.84	40.94	43.42	34.55	39.45	34.12	35.89	22.56	33.66
	Cutoff 1 SD	180.40	207.50	159.10	178.84	155.83	164.05	153.83	152.55	165.72	152.18	186.64	185.92	210.80	190.73	219.35	222.81	243.82	223.69
	Cutoff 2 SD	138.85	171.66	113.18	136.35	109.50	119.00	108.37	104.83	118.90	101.34	145.70	142.5	176.25	151.28	185.23	186.92	221.26	190.03
Approach 3																			
ROC-curve	AUC 1 SD	0.48	0.65	0.59	0.59	0.72	0.74	0.71	0.78	0.7	0.75	0.85	0.75	0.76	0.71	0.73	0.73	0.7	0.69
	Cutoff 1 SD	–	–	–	–	211.25	236.5	202.5	182.5	206	205.25	187.5	238	235.5	213	251.93	267	267	–
	AUC 2 SD	0.48	0.69	0.6	0.83	–	–	0.86	0.98	0.85	1	0.97	0.96	0.96	0.99	0.56	1	0.62	–
	Cutoff 2 SD	–	–	–	182	–	–	111.25	117.81	163	81	155.92	191.5	187.5	125	–	114	–	–
	N	119	114	100	85	43	40	62	43	73	60	49	37	52	72	72	79	70	74

**Table 4**

Means and standard deviations for all Bayley-III-NL subtests per ASQ-3 age-version.

		ASQ-3 age-versions																	
		2	4	6	8	9	10	12	14	16	18	20	22	24	27	30	33	36	42
Bayley-III-NL subtests																			
Cognition	Mean	8.97	9.06	11.72	10.49	9.77	9.00	9.39	9.26	9.88	10.28	10.98	10.16	10.27	10.83	10.74	9.99	10.24	10.41
	SD	2.58	3.09	3.11	2.46	2.52	2.63	2.88	2.75	2.91	2.94	2.64	3.25	2.71	2.72	2.83	2.84	2.74	2.85
Fine Motor	Mean	8.27	9.17	11.89	10.71	10.42	9.55	9.37	9.72	10.29	10.65	10.82	11.11	10.73	10.35	10.10	9.70	10.34	10.16
	SD	3.98	3.47	2.99	2.43	1.84	1.84	1.65	1.70	2.12	2.60	1.98	2.33	2.46	2.72	3.15	3.55	3.17	3.54
Gross Motor	Mean	9.73	10.70	10.10	10.22	9.44	9.85	9.45	9.84	10.25	9.82	10.59	10.68	10.40	9.68	10.43	10.34	10.17	10.15
	SD	3.41	2.19	2.97	2.81	3.20	3.42	3.26	2.44	3.07	3.02	2.40	2.76	2.52	2.89	3.25	2.81	2.80	2.92
Receptive Communication	Mean	10.13	9.80	10.07	10.38	10.12	10.58	9.98	9.58	10.78	10.22	11.33	10.84	10.65	10.39	10.65	10.37	10.49	10.64
	SD	3.18	2.86	2.57	2.63	2.36	2.46	2.99	2.67	3.32	3.68	2.62	3.56	3.35	2.54	2.66	2.64	2.27	2.71
Expressive Communication	Mean	10.50	10.10	9.29	10.54	9.58	10.33	10.55	9.88	11.14	10.25	10.63	10.16	10.65	10.65	10.38	10.57	10.56	10.46
	SD	2.73	2.60	3.33	2.98	2.93	2.69	3.05	3.11	2.99	2.99	2.13	3.18	2.99	2.72	2.97	2.89	2.53	2.26
	N	119	114	100	85	43	40	62	43	73	60	49	37	52	72	72	79	70	74
	N ≤ 1 SD	74	54	33	17	18	19	24	14	23	19	6	9	10	17	16	28	21	24
	N ≤ 2 SD	21	7	4	1	0	0	3	1	1	2	1	3	1	1	2	1	2	0

Note: the age-versions of the ASQ-3 that were used as basis for this study, do not match exactly with the 17 age-groups of the Bayley-III-NL that are used for administration of the assessment.

point of 2 SD on at least one Bayley-III-NL subtest. Thus, using the total ASQ-3 score—opposed to using the scores per subdomain—led to the best sensitivity and specificity. This could be explained by the fact that when using the sum of scores over all 30 items, the impact of each single item is decreased: the lack of one particular developmental skill can be compensated by the presence of 29 other skills. This is not the case when risk for developmental delay is determined using separate developmental domains which are each assessed by only 6 items; the lack of only one skill can then already lead to the conclusion that a child is at risk for a developmental delay in a certain developmental domain. Indeed, we found fewer children identified as at risk using the total ASQ-3 score than when using the separate developmental domains.

Another finding we came across in this study is that the percentages of children at risk for developmental delay according to the ASQ-3 that we found, are higher than the expected 2.3% or 15%. This discrepancy can be explained by the combination of domains and subtest results we used in our comparison methods. When the results were compared

per domain or subtest, the percentages of children with a (risk for a) delay would be more comparable to the expected percentages based on the normal distribution.

Another explanation for finding the best results using the total ASQ-3 score is that, despite similar labels, the developmental domains of the ASQ-3 and the subtests of the Bayley III-NL do not seem to measure the same constructs. A closer look at the questionnaire and test items revealed that items of the problem-solving domain of the ASQ-3 were similar to some of the items in the fine motor scale of the Bayley-III-NL; for instance, the item 'passing toy from one hand to the other'. Also, items of the fine motor scale of the ASQ-3 were similar to items in the cognitive scale of the Bayley-III-NL, like 'screwing lids on and off bottles or jars'. It is not surprising that when based on non-matching domains of development, the ASQ-3 and the Bayley-III-NL do not identify the same children as developmentally delayed. By using a total ASQ-3 score, the placement of the items does not matter anymore, indeed resulting in better correspondence between the

**Table 5**

Test characteristics of the ASQ-3 compared to the Bayley-III-NL for the first 3 comparison methods.

Comparison method	Age-versions	ASQ-3 cutoff	Bayley-III-NL cutoff	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Risk of delay ASQ-3		Delay Bayley-III-NL		(Risk of) delay on both ASQ-3 and Bayley-III-NL	
								N	%	N	%	N	%
1	2–16	1 SD	1 SD	57.2	59.3	49.1	66.9	322	47.7	276	40.6	158	57.2
		1 SD	2 SD	57.9	53.2	6.8	95.4	322	47.7	38	5.6	22	57.9
		2 SD	1 SD	19.2	89.8	65.4	61.9	94	13.8	276	40.6	53	19.2
		2 SD	2 SD	23.7	86.7	9.6	95.0	94	13.8	38	5.6	9	23.7
		1 SD	1 SD	62.0	70.4	43.1	83.7	216	38.2	150	26.5	93	62.0
	18–42	1 SD	2 SD	76.9	62.7	4.6	99.1	216	38.2	13	2.3	10	76.9
		2 SD	1 SD	27.3	89.9	49.4	77.4	83	14.7	150	26.5	41	27.3
		2 SD	2 SD	61.5	84.4	9.6	99.0	83	14.7	13	2.3	8	61.5
		2 × 1SD + 1 × 2SD	2 × 1SD + 1 × 2SD	38.0	87.6	21.3	87.6	178	26.2	100	14.7	38	38.0
		2 × 1SD + 1 × 2SD	2 × 1SD + 1 × 2SD	62.7	79.4	23.2	95.6	138	24.4	51	9.0	32	62.7
2	2–16	1 SD	1 SD	33.7	82.1	56.4	64.4	165	24.3	276	40.6	93	33.7
		1 SD	2 SD	34.2	76.3	7.9	95.1	165	24.3	38	5.6	13	34.2
		2 SD	1 SD	7.2	95.5	52.6	60.1	38	5.6	276	40.6	20	7.2
		2 SD	2 SD	10.5	94.7	10.5	94.7	38	5.6	38	5.6	4	10.5
		1 SD	1 SD	21.3	95.4	62.7	77	51	9.0	150	26.5	32	21.3
	18–42	1 SD	2 SD	69.2	92.4	17.6	99.2	51	9.0	13	2.3	9	69.2
		2 SD	1 SD	7.3	99.3	78.6	74.8	14	2.5	150	26.5	11	7.3
		2 SD	2 SD	38.5	98.4	35.7	98.5	14	2.5	13	2.3	5	38.5
		2 × 1SD + 1 × 2SD	2 × 1SD + 1 × 2SD	38.0	87.6	21.3	87.6	178	26.2	100	14.7	38	38.0
		2 × 1SD + 1 × 2SD	2 × 1SD + 1 × 2SD	62.7	79.4	23.2	95.6	138	24.4	51	9.0	32	62.7

Note: NPV = negative predictive value; PPV = positive predictive value.

A cutoff point of 1 SD below the mean reflects a score that is 1 SD or more below the mean and a cutoff point of 2 SD below the mean reflects a score that is 2 SD or more below the mean. Method 1: cutoff points are 1 or 2 SD below the mean on at least one domain or subtest.

Method 2: cutoff points are 1 SD below the mean on at least two domains/subtests and/or 2 SD below the mean on at least one domain/subtest.

Method 3: cutoff points are based upon 1 or 2 SD below the mean of the total score for all domains of the ASQ-3 and below 1 or 2 SD below the mean on at least 1 subtest of the Bayley-III-NL.

results of the ASQ-3 and Bayley-III-NL. However, the agreement between both instruments is still not sufficient as is shown by the insufficient sensitivity value. Therefore, a critical evaluation is necessary whether both instruments truly measure the same.

Regarding all three comparison methods, the best sensitivity and specificity values were found for the oldest age-cohort. The finding of better sensitivity and specificity values for the older age-cohort in comparison to the younger age-cohort is in line with the findings of Simard et al. [22] and Schonhaut et al. [23]. They also found that with increasing age, sensitivity and specificity values increased. Thus, it seems that developmental delays are easier to identify when children get older. This could be explained by the fact that the variation in ages at which children reach certain developmental milestones becomes smaller with age [31]. Consequently, developmental delay might be more apparent when children are older at age of assessment and therefore easier for both parents and professionals to evaluate.

The finding of better sensitivity and specificity in the older age cohort could also indicate that the items per age-version and the ages covered per age-version of the ASQ-3 are more adequate for older children than younger children. Earlier studies indicated that Dutch children show a different development pathway than US children and that the magnitude of these differences depends on age and developmental domain [15,16] with differences becoming smaller as the age of the child increases [15]. This indicates that, especially for the younger age cohort, items within a certain ASQ-3 age version might be too easy or too difficult for Dutch children. It could also be that, because of the developmental pace of Dutch children, the age-window per age-version is not completely suitable for Dutch children. This is supported by the large variation between the means and standard deviations of the ASQ-3 age-versions, especially for the younger children, sometimes even resulting in a negative cutoff point using 2 SD. Further research is needed to develop adequate Dutch ASQ-3 age-versions.

Another reason for the moderate agreement between the results of the ASQ-3 and the Bayley-III-NL might be that children can show different behavior and skills when they are in a strange situation such as a Bayley-III-NL assessment. Consequently, parents might observe different skills at home than the skills the child showed during the assessment, causing differences in the scores of the ASQ-3 and the Bayley-III-NL which might have resulted in a moderate agreement between the results of the two instruments.

Another finding we came across was that of the children with missing data, 33% had a developmental delay according to the Bayley-III-NL. This indicates that completing the ASQ-3 might be difficult for parents of children with a developmental delay. Future research should focus on why parents with a delayed child would find the ASQ-3 difficult to complete.

A major strength of this study is that the sample is a large community sample with a broad age-range. Many previous studies on the ASQ-3 only included children with a risk of a developmental delay and as such can only provide limited information on the adequacy of the ASQ-3 in the way it is often used: Screening overall healthy children in the community. The current study overcomes this by including a large, representative community sample. In addition, most previous studies focused on only one or a few age-versions of the ASQ-3. Given of the broad age-range of the sample in this study, most age-versions of the ASQ-3 could be evaluated in this study.

A disadvantage of using a community sample is that only a small number of children had a developmental delay. Therefore, cutoff points based on ROC curves could not be determined for all age-version as not for every ASQ-3 age-version children are delayed according to the Bayley-III-NL. This raises questions concerning the quality of the other cutoff points that were determined in this study, which might have led to the moderate agreement between the results of the ASQ-3 and the Bayley-III-NL.

In conclusion, most Dutch infants and toddlers that showed age appropriate functioning according to the ASQ-3, also showed age

appropriate developmental outcome according to the Bayley-III-NL. However, the sensitivity of the ASQ-3 was found to be only low to moderate. It is concluded that the 18–42 months version of the ASQ-3 in the Netherlands might have the best potential to use as a screenings instrument, but only when a total score for the ASQ-3 is calculated and 1 SD below the mean is used as cutoff point.

## Conflict of interest statement

None declared.

## Acknowledgments

This research was supported by ZonMW project number 15071.3001. The authors wish to thank all children and parents that participated in this study, Endy den Exter Blokland for the coordination of the data collection, and all test assistants that conducted the Bayley-III-NL.

## References

- [1] Reynolds AJ, Temple JA, Ou S, Robertson DL, Mersky JP, Topitzes JW, et al. Effects of a school-based, early childhood intervention on adult health and well-being: a 19-year follow-up of low-income families. *Arch Pediatr Adolesc Med* 2007;161(8):730–9.
- [2] Reynolds AJ, Temple JA, Robertson DL, Mann EA. Long-term effects of an early childhood intervention on educational achievement and juvenile arrest. *JAMA* 2001;285(18):2339–46.
- [3] King TM, Glascoe FP. Developmental surveillance of infants and young children in pediatric primary care. *Curr Opin Pediatr* 2003;15(6):624–9.
- [4] Glascoe FP, Dworkin PH. The role of parents in the detection of developmental and behavioral problems. *Pediatrics* 1995;95(6):829.
- [5] Glascoe FP. Parents' evaluation of developmental status. Nashville, TN: Ellsworth & Vandermeer Press, Ltd; 1997.
- [6] Glascoe FP. Using parents' concerns to detect and address developmental and behavioral problems. *J Spec Pediatr Nurs* 1999;4(1):24–35.
- [7] Glascoe FP. Parents' evaluation of developmental status: how well do parents concerns identify children with behavioral and emotional problems? *Clin Pediatr* 2003;42(2):133–8.
- [8] Hix-Small H, Marks K, Squires J, Nickel R. Impact of implementing developmental screening at 12 and 24 months in a pediatric practice. *Pediatrics* 2007;120(2):381.
- [9] Squires J, Twombly E, Bricker D, Potter L. The ASQ-3 user's guide. 3rd ed. Baltimore, MD: Brookes; 2009.
- [10] Van Baar AL, Steenis LJP, Verhoeven M, Hessen D. Bayley-III-NL; technische handleiding. Amsterdam, the Netherlands: Pearson Assessment and Information B.V; 2014.
- [11] Glascoe FP, Foster EM. An economic analysis of developmental detection methods. *Pediatrics* 1997;99(6):830 [06].
- [12] Squires J, Bricker D. The ASQ-2 user's guide. 2nd ed. Baltimore, MD: Paul H. Brookes; 1999.
- [13] Bayley N. The Bayley Scales of Infant Development. 2nd ed. San Antonio, TX: The Psychological Corporation; 1993.
- [14] Bayley N. The Bayley Scales of Infant and Toddler Development. San Antonio, TX: Psychological Corporation; 2006.
- [15] Steenis LJP, Verhoeven M, Hessen D, Van Baar AL. Performance of Dutch children on the Bayley III: a comparison study of Dutch and US norms; 2015[in preparation].
- [16] Westera JJ, Houtzager BA, Overdiek B, Van Wassenar AG. Applying Dutch and US versions of the BSID-II in Dutch children born preterm leads to different outcomes. *Dev Med Child Neurol* 2008;50(6):445–9.
- [17] Chinta S, Walker K, Halliday R, Loughran-Fowlds A, Badawi N. A comparison of the performance of healthy Australian 3-year-olds with the standardized norms of the Bayley Scales of Infant and Toddler Development (version-III). *Arch Dis Child* 2014;99(7):621–4.
- [18] Walker K, Badawi N, Halliday R, Laing S. Brief Report: performance of Australian children at one year of age on the Bayley Scales of Infant and Toddler Development (version III). *Aust Educ Dev Psychol* 2010;27(1):54–8.
- [19] Yu Y, Hsieh W, Hsu C, Chen L, Lee W, Chiu N, et al. A psychometric study of the Bayley Scales of Infant and Toddler Development—3rd edition for term and preterm Taiwanese infants. *Res Dev Disabil* 2013;34(11):3875–83.
- [20] Bian X, Yao G, Squires J, Hoselton R, Chen C, Murphy K, et al. Translation and use of parent-completed developmental screening test in Shanghai. *J Early Child Res* 2012;10(2):162–75.
- [21] Gollenberg AL, Lynch CD, Jackson LW, McGuinness BM, Msall ME. Concurrent validity of the parent-completed Ages and Stages Questionnaires, with the Bayley Scales of Infant Development II in a low-risk sample. *Child Care Health Dev* 2010;36(4):485–90.
- [22] Simard MN, Luu TM, Gosselin J. Concurrent validity of ages and stages questionnaires in preterm infants. *Pediatrics* 2012;130(1):e108–14.
- [23] Schonhaut I, Armijo I, Schonstedt M, Alvarez J, Cordero M. Validity of the ages and stages questionnaires in term and preterm infants. *Pediatrics* May 2013;131(5):e1468–74.



- [24] Skellern CY, Rogers Y, O'Callaghan MJ. A parent-completed developmental questionnaire: follow up of ex-premature infants. *J Paediatr Child Health* 2001;37(2):125–9.
- [25] Smith C, Wallen M, Walker K, Bundy A, Rolinson R, Badawi N. Validity of the fine motor area of the 12-month Ages and Stages Questionnaire in infants following major surgery. *Phys Occup Ther Pediatr* 2012;32(3):260–71.
- [26] Woodward BJ, Papile L, Lowe JR, Laadt VL, Shaffer ML, Montman R, et al. Use of the Ages and Stages Questionnaire and Bayley Scales of Infant Development-II in neurodevelopmental follow-up of extremely low birth weight infants. *J Perinatol* 2011;31(10):641–6.
- [27] Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods* 2002;7(2):147.
- [28] Swets JA. Measuring the accuracy of diagnostic systems. *Science* Jun 3 1988; 240(4857):1285–93.
- [29] Horwitz S, Irwin JR, Briggs-Gowan MJ, Bosson Heenan JM, Mendoza J, Carter AS. Language delay in a community cohort of young children. *J Am Acad Child Adolesc Psychiatry* 2003;42(8):932–40.
- [30] Sustersic B, Paro-Panjan D. Assessment of general movements in relation to neurologic signs at age two years. *Pediatr Neurol* 2008;39(2):108–12.
- [31] Losardo A, Notari-Syverson A, Bricker DD. Alternative approaches to assessing young children. Paul H. Brookes Publishing Company; 2001.
- [32] CBS Statline. Available at: [http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLNL&PA=37296ned&D1=a&D2=0,10,20,30,40,50,60,\(1-1\),I&HD=130605-0924&HDR=G1&STB=T](http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLNL&PA=37296ned&D1=a&D2=0,10,20,30,40,50,60,(1-1),I&HD=130605-0924&HDR=G1&STB=T); 2013. [Accessed 05/14, 2014].