



Assessing social-emotional development in infants and toddlers using parent-reports: Comparing the ASQ-SE-NL to the Social-Emotional Scale of the Bayley-III-NL

Lisa J.G. Krijnen^{*}, Marjolein Verhoeven, Anneloes L. van Baar

Child and Adolescent Studies, Utrecht University, Utrecht, the Netherlands

ARTICLE INFO

Keywords:

Social-emotional development
Ages and Stages Questionnaire Social-Emotional
Bayley-III
Infants
Toddlers
Community sample

ABSTRACT

Background: The Ages and Stages Questionnaire Social-Emotional (ASQ-SE) is a parent-report screening instrument designed to assess children's social-emotional development and detect those at risk for delay or problems. Psychometric properties of this questionnaire have been studied before, but the ASQ-SE has never been compared to the Social-Emotional Scale of the Bayley-III (Bayley-III-SE).

Aim: To compare the Dutch ASQ-SE (ASQ-SE-NL) to the Dutch Bayley-III-SE (Bayley-III-NL-SE; criterion measure).

Method: A Dutch community sample of mothers with children aged 3–41 months ($n = 1014$) filled out both questionnaires. Cut-off scores for the ASQ-SE-NL were determined using: 1) >1 SD above the mean and 2) ROC curves. For the Bayley-III-NL, Dutch norm scores were used.

Results: Specificity (70.8% and 88.5%) and screen-out accuracy (0.65 and 0.77) of the ASQ-SE-NL were good. Sensitivity was only sufficient (70.6%) when using ROC curves and only for the ASQ-SE-NL ≥ 18 months age versions. Screen-in accuracy was insufficient (<0.49). Positive predictive value was 34.7% and 32.7%, and negative predictive value was 87.5% and 92.3%. False positive cases on the ASQ-SE-NL scored significantly lower on the Bayley-III-NL-SE than true negative cases.

Conclusion: Using the Bayley-III-NL-SE as the criterion, the ASQ-SE-NL performed well in identifying children not at risk for delay or problems. The ASQ-SE-NL sufficiently detected children at risk for delay or problems in the ≥ 18 months ASQ-SE-NL age versions when cut-off scores were determined by ROC curves. The ASQ-SE-NL can be used in a monitoring routine, but early rescreening is advised after a positive test result, given the number of false positive results.

1. Introduction

Early detection of developmental problems or delay in children is crucial to provide timely care for enhancing outcomes, including mental health [1–3]. Infants and toddlers develop at an outstandingly rapid pace [4], underlining the need for adequate screening instruments. One of the fast-evolving domains concerns social-emotional development, which consists of the ability to experience, express and regulate emotions in an age-appropriate manner, to develop and maintain healthy relationships with others, and to feel confident to explore the environment and learn [5–7].

Multiple parent-report instruments have been developed to assess children's social-emotional development, as parents have been found to form reliable and cost-efficient sources to report about the child's development [8–10]. Different types of instruments exist. A screening questionnaire is often used as a first step to briefly evaluate children and identify those at risk for delay or problems who might need further assessment [11]. The Ages and Stages Questionnaire Social-Emotional (ASQ-SE, [12]) is such a screening questionnaire that aims to evaluate different dimensions of social-emotional development and to identify children at risk for delays or problems. If a child scores at risk, further assessment is needed to investigate more extensively how the child

Abbreviations: ASQ-SE-NL, the Dutch translation of the Ages and Stages Questionnaire - Social-Emotional; AUC, Area Under Curve; Bayley-III-NL-SE, the Dutch translation of the Social-Emotional Scale of the Bayley-III; PPV, positive predictive value; NPV, negative predictive value; CUI, Clinical Utility Index; UI+, screen-in accuracy; UI−, screen-out accuracy; ROC, Receiver Operating Characteristic.

^{*} Corresponding author at: Room E2.44, Heidelberglaan 1, 3584 CS Utrecht, the Netherlands.

E-mail addresses: l.j.g.krijnen@uu.nl (L.J.G. Krijnen), j.c.t.verhoeven@uu.nl (M. Verhoeven), a.l.vanbaar@uu.nl (A.L. van Baar).

<https://doi.org/10.1016/j.earlhumdev.2021.105439>

Received 11 May 2021; Received in revised form 20 July 2021; Accepted 30 July 2021

Available online 8 August 2021

0378-3782/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

performs compared to same-aged children. The Bayley-III is a developmental test that can be used for this aim regarding cognition, motor and language development and is often considered as the gold standard for assessing the development of young children for both practice and research purposes [13,14]. The Bayley-III also contains a Social-Emotional Scale which consists of a parent-report questionnaire (Bayley-III-SE; [15]). We are studying if the Dutch versions of the ASQ-SE and the Bayley-III-SE can be used in conjunction in a monitoring routine. Both questionnaires aim to identify children with – or at risk for – developmental problems or delay, but also provide more information regarding different dimensions of social-emotional development that may show specific difficulties or advancements. To our knowledge, this is the first study comparing these instruments. In the Netherlands, two studies have investigated the psychometric properties of the Dutch ASQ-SE (ASQ-SE-NL) [16,17]. Specificity and sensitivity were calculated using the Child Behavior Checklist as the criterion (CBCL; [18]), a parent-report questionnaire focused on the presence of rather concrete behavior problems — a somewhat different concept than social-emotional development. Specificity was always >0.90 due to their methods. Results suggested lower psychometric properties for the 6 and 12 months ASQ-SE-NL age version (sensitivity: 0.28–0.38, correlation to CBCL: 0.32–0.35) [16] than in the older, 24, 36, and 48 age versions (sensitivity: 0.63–0.66, correlation to CBCL: 0.54–0.63) [16,17]. However, the CBCL has been designed for assessment of children aged 18–66 months and was therefore answered months later than the ASQ-SE in the 6–12 months age groups [16]. Due to the age range of the CBCL and the concept measured, the CBCL may not form an ideal criterion measure. A measure that may better fit the construct that the ASQ-SE aims to assess and covers a large part of its age range is the Bayley-III-SE. Moreover, population norms for the Dutch Bayley-III-SE (Bayley-III-NL-SE) have been provided for the Netherlands [19].

The current study will compare the ASQ-SE-NL to the Bayley-III-NL-SE in a Dutch community sample. Using the Bayley-III-NL-SE as a criterion, sensitivity, specificity, positive predictive value and negative predictive value of the ASQ-SE-NL will be calculated in addition to the clinical utility. Furthermore, for children with false negative scores or

false positive scores on the ASQ-SE-NL, mean scores on the Bayley-III-NL-SE will be explored as these children may not be on the same developmental level as true positives or true negatives respectively [20,21].

2. Methods

2.1. Participants

Parents of children aged 2–42 months participated in a larger study focused on establishing Dutch norms for the Dutch Bayley-III (Bayley-III-NL) [19,22–24]. A sample consisting of Dutch children aged 0–3.5 years old whose mothers answered the ASQ-SE-NL and the Bayley-III-NL-SE is used for the current study ($n = 1014$). The sample consisted of healthy children (93.9%), and prematurely born children (4.3%) or children with other risk factors (1.8%) known to hinder the development, such as cerebral palsy, tracheomalacia, or Williams syndrome. See Table 1 for an overview of the participants' characteristics.

2.2. Procedure

Parents of children were recruited through day-care centers, child health care centers, advertisements in newspapers and personal connections. At recruitment parents were asked to answer questions regarding birth weight, health and gestational age of the child to determine which age version of the ASQ-SE-NL and Bayley-III-NL-SE had to be filled out. Preterm children (gestational age < 37 weeks) received the age version that matched their age corrected for the number of weeks they were born prematurely. Parents who were interested in participation received an informed consent form and the questionnaires, by mail. After two weeks, parents handed in the questionnaires during their appointment for the administration of the Bayley-III-NL. The Medical Ethical Committee of Utrecht University Medical Center approved this study (NL29428.041.09).

Table 1
Participants' characteristics for the total sample and per ASQ-SE-NL age version.

ASQ-SE-NL age version	Total ($n = 1014$)	6 ($n = 247$)	12 ($n = 159$)	18 ($n = 162$)	24 ($n = 119$)	30 ($n = 135$)	36 ($n = 192$)
Age in months ^a							
Mean (SD)	19.54 (11.76)	5.25 (1.69)	11.13 (1.66)	17.36 (1.69)	23.55 (1.66)	29.52 (1.62)	36.98 (2.67)
Range	3.00–41.00	3.00–8.00	9.00–14.00	15.00–20.00	21.00–26.00	27.00–32.00	33.00–41.00
Gender							
Male, n (%)	514 (50.7%)	136 (55.1%)	84 (52.8%)	71 (43.8%)	53 (44.5%)	67 (49.6%)	103 (53.6%)
Female, n (%)	500 (49.3%)	111 (44.9%)	75 (47.2%)	91 (56.2%)	66 (55.5%)	68 (50.4%)	89 (46.4%)
Gestational age in weeks, mean (SD) ^b	39.59 (1.63)	39.68 (1.44)	39.59 (1.48)	39.78 (1.75)	39.75 (1.41)	39.80 (1.61)	39.40 (1.98)
Health							
No known risk factor, n (%)	952 (93.9%)	237 (96.0%)	150 (94.3%)	147 (90.7%)	114 (95.8%)	127 (94.1%)	177 (92.2%)
Premature birth, n (%) ^b	44 (4.3%)	8 (3.2%)	5 (3.1%)	11 (6.8%)	4 (3.4%)	4 (3.0%)	12 (6.2%)
Other known risk factor, n (%)	18 (1.8%)	2 (0.8%)	4 (2.5%)	4 (2.5%)	1 (0.8%)	4 (3.0%)	3 (1.6%)
Ethnic background mother							
Dutch n (%) ^c	908 (89.5%)	218 (88.3%)	141 (88.7%)	52 (93.8%)	106 (89.1%)	118 (87.4%)	173 (90.1%)
Educational level mother ^d							
Low, n (%)	107 (10.6%)	30 (12.1%)	22 (13.8%)	12 (7.4%)	20 (16.8%)	17 (12.6%)	19 (9.9%)
Intermediate, n (%)	392 (38.7%)	92 (37.2%)	64 (40.3%)	63 (38.9%)	39 (33.6%)	48 (35.6%)	70 (36.5%)
High, n (%)	507 (50.0%)	122 (49.4%)	73 (45.9%)	87 (53.7%)	59 (49.6%)	70 (41.8%)	103 (53.6%)

Note: ASQ-SE-NL = the Dutch translation of the Ages and Stages Questionnaire – Social-Emotional; SD = Standard Deviation. Aside from age in months, none of the demographic characteristics differed significantly between the ASQ-SE-NL age versions.

^a Age was rounded down to the nearest month and for preterm children (gestational age < 37 weeks) their corrected age for prematurity is shown.

^b Gestational age was unknown for seven participants. Three participants were born extremely preterm, gestational age < 32.

^c For nine participants, the ethnicity of the mother was unknown.

^d Low educational level refers to special education, (less than) primary school or pre-vocational secondary education and lower secondary education; intermediate educational level refers to upper secondary and post-secondary non-tertiary education, such as pre-university education or secondary vocational education; high educational level refers to higher professional education or university bachelor/master/doctoral or equivalent education. For eight participants (0.8%), the educational level of the mother was unknown. This classification is based on the 'Standaard Onderwijsindeling 2021' [25].

2.3. Instruments

2.3.1. Ages and Stages Questionnaire Social-Emotional (ASQ-SE; [12])

The ASQ-SE assesses children's social-emotional development and screens for delay or problems by addressing seven behavioral dimensions: self-regulation, compliance, communication, adaptive functioning, autonomy, affect, and interaction with people [12]. The ASQ-SE can be administered in children aged from 3 to 66 months old, using the age version closest to the child's age: 6, 12, 18, 24, 30, 36, 48 and 60 months. Given the age range in our sample, the 6–36 age versions were used. Every questionnaire contains 19–33 scored items and can be answered by the primary caregiver, which takes 10–15 min. Parents indicate whether the child shows the described behavior 'most of the time' (0 points), 'sometimes' (5 points) or 'rarely/never' (10 points). Additionally, the parent/caregiver can mark for each question whether s/he has worries about the described behavior, leading to 5 additional points per question. Sum scores are calculated by adding points from all items and the expressed worries. A higher score reflected a lower level of social-emotional development. We translated the ASQ-SE to Dutch: ASQ-SE-NL. As there are no norms for the ASQ-SE-NL, we calculated cut-off scores for the current sample to determine when a child was 'at risk'.

2.3.2. Bayley-III Social-Emotional Scale (Bayley-III-SE; [15,19])

The Bayley-III-SE concerns a parent-report questionnaire to assess functional emotional milestones of children aged 0–42 months, in the following dimensions: self-regulation, interest in the world, engagement in relationships, use of emotions in a meaningful way, use of interactive gestures or emotional signals to communicate and solve problems, use of ideas or symbols to express intentions, feelings and more than solely basic needs, and the ability to create bridges between ideas and emotions [15,26]. Within these dimensions, questions regarding sensory processing skills (i.e. how the individual reacts to sensations such as touch, smell, sight, taste, movement and sound) are also included. The questionnaire contains 11 questions for the youngest age group (0–3 months old) and questions are added as the child gets older with a maximum of 35 questions. Answer options ranged from 'can't tell' (0 points), 'none of the time' (1 point), 'some of the time' (2 points), 'half of the time' (3 points), 'most of the time' (4 points), 'all of the time' (5 points). Adding the points leads to a sum score with a higher score reflecting a better social-emotional development.

The Bayley-4 has recently been developed, but the items of the Social-Emotional Scale did not change from the 3rd to the 4th edition [27]. The Bayley-III-SE has been translated to Dutch and norms have been created using a representative sample of 1845 mothers of which we used a subsample for the current study [19]. Results showed a good item-rest correlation [28]. Reliability and ability to discriminate between children at risk for developmental delay or problems versus typically developing children was good. Day norms, reflecting norms for age in days, instead of an age range, were created to ensure the best possible reference scores. Standardized scores with mean = 100 and 1 SD = 15 and a range of 40–160 were used. A child was considered having developmental problems or delay when scoring below an index score of 85 (1 SD below the standardized mean of 100), referring to the 16th percentile.

2.4. Statistical analyses

SPSS version 25 was used to analyze the data [29]. Missing data of the ASQ-SE-NL was handled using the ASQ-SE manuals' guidelines [12]. As there are no Dutch norms for the ASQ-SE-NL, two approaches were used to determine cut-off scores. For Approach 1, the conventional approach of 1 SD above the mean as cut-off criterion was used per age version in this sample. However, in a healthy population most children do not have developmental delay or problems, resulting in a positively skewed distribution for the scores on the ASQ-SE. Therefore, for Approach 2, empirical ROC curves were computed to determine the

most optimal cut-off scores as this approach does not require a normal distribution [30]. ROC curves were computed for each ASQ-SE-NL age version, using sum scores on the ASQ-SE-NL and a dichotomous variable of the Bayley-III-NL-SE as the criterion (reflecting no delay or problems versus having delay or problems). We used the Youden's Index to determine cut-off scores [31]. The Area Under Curve (AUC) was calculated with an AUC of 0.50–0.70 reflecting low accuracy, 0.70–0.90 moderate accuracy and 0.90–1.00 high accuracy [32]. If AUC scores were <0.70, cut-off scores could not be determined as the accuracy would be low.

For both approaches, children were classified into one of the following conditions: 1) true positive – i.e. (risk for) delay or problems on both questionnaires, 2) true negative – no (risk for) delay or problems on both questionnaires, 3) false positives – i.e. at risk on the ASQ-SE-NL, but *no* delay or problems on the Bayley-III-NL-SE, and 4) false negatives – i.e. scoring *not* at risk on the ASQ-SE-NL but showing delay or problems on the Bayley-III-NL-SE. First, with independent samples *t*-tests we tested whether children with false positive and false negative scores scored differently on the Bayley-III-NL-SE than true negative and true positive cases respectively. Hence, specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV) were calculated. These calculations were first carried out per ASQ-SE-NL age group. Next, the classifications in terms of true-/false positive/negative cases of all age groups were grouped together to calculate these measures for the total sample.

For sensitivity and specificity percentages $\geq 70\%$ are considered good [21,22,33]. There are no clear guidelines to interpret the PPV and NPV because these values strongly depend on the prevalence of the condition/disease in the sample [34]. To illustrate: if the prevalence of a disease is 50% in a sample of 1000 children, 500 children are diseased and 500 children are healthy. If the sensitivity of the given screening instrument is 70%, there are 350 true positives and 150 false positives resulting in a PPV of 70%. However, when the prevalence decreases to 15%, there are 150 diseased cases and 850 healthy cases, with 105 true positive cases and 255 false positive cases when sensitivity is 70%. This results in a PPV of 29.2% which corresponds to a larger number of false positives compared to the former example. To provide a clearer meaning to these measures, we calculated the Clinical Utility Index (CUI; [35–37]) which gives an indication of the clinical utility of a positive screening result, i.e. the screen-in accuracy (UI+) and negative test result, i.e. the screen-out accuracy (UI–). The CUI takes into account both the occurrence of the condition (i.e. sensitivity, specificity) as the discriminative ability of the screening instrument (i.e. PPV, NPV) [38]. As both the PPV and sensitivity provide information about the accuracy of a positive screening result, and the NPV and specificity about a negative test result, it is useful to combine these measures to get more insight into the accuracy of a positive/negative test result. See Table 2 for the interpretation and calculations of the measures.

3. Results

3.1. Cut-off scores

3.1.1. Bayley-III-NL-SE

On the Bayley-III-NL-SE, 15.8% of the children scored as having developmental delay or problems. For ROC curves, this percentage was different due to the exclusion of the 6 and 12 months age versions of the ASQ-SE-NL: the AUC for these age versions was below 0.70. Cut-off scores could therefore not be determined (see Table 3). In the subsample of the ASQ-SE-NL ≥ 18 age versions, 16.8% of the children scored delayed or as having problems on the Bayley-III-NL-SE.

3.1.2. ASQ-SE-NL

Table 3 shows the cut-off scores and the number of children that scored above the cut-off of 1 SD per ASQ-SE-NL age version. The percentage of children identified as 'at risk' ranged from 12.6% to 19.8%.

Table 2
Calculations and classifications of the measurements.

Measure	Calculation	Classification
Sensitivity	True positives / (true positives + false negatives)	≥70% good
Specificity	True negatives / (true negatives + false positives)	≥70% good
PPV	True positives / (true positives + false positives)	n.a.
NPV	True negatives / (true negatives + false negatives)	n.a.
UI+	Sensitivity * PPV	Excellent utility ≥0.81, good utility ≥0.64 and satisfactory utility ≥0.49 and poor utility <0.49
UI−	Specificity * NPV	Excellent utility ≥0.81, good utility ≥0.64 and satisfactory utility ≥0.49 and poor utility <0.49

Note: PPV = positive predictive value; NPV = negative predictive value; n.a. = not applicable; UI+ = “screen-in accuracy”, UI− = “screen-out accuracy”.

The percentage of children scoring above the cut-off score on the ASQ-SE-NL using ROC curves, varied from 32.3% to 40.7%. See Table 3.

3.2. Comparison of ASQ-SE-NL and Bayley-III-NL-SE

The results showed no consistent difference for the different age versions. Results are presented grouping all ASQ-SE-NL age versions.

3.2.1. Comparison between classifications

Independent t-tests showed that false positive cases scored approximately three to four index points lower on the Bayley-III-NL-SE than true negative cases for both approaches. True positive cases did not score significantly different from false negative cases. See Table 4.

3.2.2. Approach 1: 1 SD > mean

Sensitivity of the ASQ-SE-NL was 32.5% and the PPV was 34.7% with a corresponding UI+ of 0.11. Specificity was 88.5%, the NPV 87.5% and the UI− score was 0.77. See Table 5.

3.2.3. Approach 2: ROC curve

Cut-off scores could only be determined for the 18 months and older ASQ-SE-NL age versions, as the AUC of the 6 and 12 months versions was too low to determine cut-off scores. For the ≥18 months age versions, the AUC was significant and showed moderate accuracy, indicating that

the prediction with the ASQ-SE was better than a random guess. Sensitivity was 70.6%, the PPV was 32.7% and the UI+ 0.23. Specificity was 70.8% and the NPV 92.3%. The UI− was 0.65. See Table 5.

4. Discussion

The current study compared the ASQ-SE-NL to the Bayley-III-NL-SE in a Dutch community sample of children between 3 and 41 months old. Compared to the Bayley-III-NL-SE, specificity of the ASQ-SE-NL was good (≥70%) indicating that a child without delay or problems according to the Bayley-III-NL-SE was likely to be classified *not at risk* (i.e. true negative and not a false positive) by the ASQ-SE. The NPV, referring to the likelihood that a child with a negative screening result indeed shows no delay or problems (i.e. not a false negative) was above 85%. The screen-out accuracy was also good, meaning that the ASQ-SE-NL showed a good clinical utility in screening out those without delay or problems. These findings held for both approaches used. The sensitivity of the ASQ-SE-NL was good when cut-off scores were based on ROC curves (70.6%). However, when cut-off scores were based on >1 SD, sensitivity was insufficient (i.e. <70%), indicating too many false negative findings. The PPV, which indicates the likelihood of a positive screening result being a true positive and not a false positive, may seem low (32.7% and 34.7%). It is, however, important to note that given the rather low prevalence of developmental delay or problems in the current sample (15.8%–16.8%), these values are not expected to be higher given the sensitivity/specificity found in this study. The screen-in accuracy was below the criterion indicating that the ASQ-SE-NL showed poor clinical utility in case-finding (i.e. correctly identifying children *at risk* for delay or problems). Furthermore, false positive cases scored significantly lower on the Bayley-III-NL-SE than true negative cases, but false negative cases did not score different on the Bayley-III-NL-SE than true positive cases. The overall results were better when using ROC curves to determine cut-off scores than >1 SD.

It should be noted that the results based on ROC curves only apply to the 18–36 months ASQ-SE-NL age versions. The AUC of the 6–12 months age versions was below 0.70 indicating low accuracy for prediction and cut-off scores were therefore not calculated, precluding the use of these versions. The poor prediction of the 6–12 months age versions seem in line with previous research showing poor performance of these ASQ-SE-NL age versions [16]. One explanation for the poor performance of these age versions may be that regulating and understanding emotions is a skill that becomes more apparent after 12 months of age. Entering toddlerhood is paired with behavioral and regulatory changes. Toddlers learn to identify and understand their own and others' internal states [39] and are rapidly developing behavioral control [39]. At the same time, aggressive behavior tends to peak in this phase [39]. More

Table 3
Cut-off points for the ASQ-SE-NL per age-version.

ASQ-SE-NL age version		6 (n = 247)	12 (n = 159)	18 (n = 162)	24 (n = 121)	30 (n = 135)	36 (n = 192)
ASQ-SE-NL score	Mean	18.13	18.13	20.89	20.61	31.23	25.98
	SD	13.66	15.17	13.87	17.34	21.92	19.63
	Range	0–70	0–99.52	0–60.00	0–115.0	0–110.00	0–110.00
Approach 1: >1 SD	Cut-off 1 SD	31.79	33.31	34.77	37.95	53.14	45.52
	n, (%)	31 (12.6%)	21 (13.2%)	32 (19.8%)	17 (14.0%)	20 (14.8%)	29 (15.1%)
Approach 2: ROC	AUC	0.61 [†]	0.65 ^{*,†}	0.75 ^{***}	0.76 ^{**}	0.72 ^{**}	0.73 ^{***}
	Cut-off	–	–	22.90	22.90	33.04	31.03
	n, (%)	–	–	66 (40.7%)	41 (34.5%)	51 (37.8%)	62 (32.3%)

Note: ASQ-SE-NL = the Dutch translation of the Ages and Stages Questionnaire — Social-Emotional; AUC = Area Under Curve; SD = standard deviation; ROC = Receiver Operating Characteristic. Scores are presented for the ASQ-SE-NL with a lower score on the ASQ-SE-NL referring to a higher social-emotional functioning.

* Significant result $p < 0.05$.

** Significant result $p < 0.01$.

*** Significant result $p < 0.001$.

[†] AUC was below 0.70, indicating a low accuracy. Cut-off scores could therefore not be determined.

Table 4

Scores on the Bayley-III-NL-SE and number of participants per classification in terms of true-/false positive/negative cases.

		Positive result on Bayley-III-NL-SE	Negative result on Bayley-III-NL-SE
Positive result on ASQ-SE-NL	Approach	<i>True positives</i>	<i>False positives</i>
	>1 SD ^a , n (%)	52 (5.1%)	98 (9.7%)
	Mean score (SD)	79.23 (6.82)	102.03 (9.47)**
	ROC ^b , n (%)	72 (11.8%)	148 (24.3%)
	Mean score (SD)	79.86 (6.66)	103.94 (10.51)*
Negative result on ASQ-SE-NL	Approach	<i>False negatives</i>	<i>True negatives</i>
	>1SD ^a , n (%)	108 (10.7%)	756 (74.6%),
	Mean score (SD)	81.02 (5.37)	106.37 (11.27)
	ROC ^b , n (%)	30 (4.9%)	358 (58.9%)
	Mean score (SD)	81.50 (4.76)	106.96 (11.27)

Note: Bayley-III-NL-SE = the Dutch translation of the Social-Emotional Scale of the Bayley-III; ASQ-SE-NL = the Dutch translation of the Ages and Stages Questionnaire – Social-Emotional; SD = Standard Deviation; ROC = Receiver Operating Characteristic.

^a n = 1014.

^b n = 608.

* $p < 0.01$. False positive cases scored lower on the Bayley-III-NL-SE than true negative cases using the approach of ROC curves.

** $p < 0.001$. False positive cases scored lower on the Bayley-III-NL-SE than true negative cases using the approach of >1 SD.

variation between toddlers in the development of these skills and behaviors emerge and can be identified by a screening instrument. Another explanation concerns differences between the two instruments: while the ASQ-SE 6 and 12 months age versions contain questions regarding problems with sleeping and eating, the Bayley-III-SE does not include these aspects and primarily focuses on sensory processing skills for the youngest children.

The differences between these questionnaires are also – to a certain extent – present in the older age versions which may explain why there is no perfect agreement between both instruments. The ASQ-SE focuses more on emotions and acting out behavior, whereas the Bayley-III-SE focuses more on sensory processing skills. Furthermore, the Bayley-III-SE uses a 5-point Likert scale, whereas the ASQ-SE provides 3 answer options. This could have led to different results, e.g. if the parent was in doubt and chose ‘never’ on the ASQ-SE, s/he could have given a slightly more optimistic answer (1 point higher than the lowest) on the Bayley-III-SE leading to a better result on the Bayley-III-SE than on the ASQ-SE.

The current findings lead to several implications for clinical practice. Our results using ROC curves lead to a relatively large number of children classified as at risk. Screening a rather healthy population, i.e. low prevalence of delay or problems, leads to a relatively high number of false positives as opposed to true positives (i.e. low PPV) compared to screening high-risk samples [40]. This is an inevitable problem when screening community samples if specificity is not 100% and it is therefore important that the involved practitioners are aware of this issue. However, we would like to point out again that children classified as false positive cases scored significantly lower on the Bayley-III-NL-SE than true negative cases. Previous research also found that children screened as false positives on developmental screening instruments scored substantially lower than true negatives on measures of intelligence, language and academic achievement [20]. Therefore these ‘false positive’ children may still form a group for whom further assessment is not an unnecessary cost, but a beneficial service to facilitate additional support [20,21]. This is also relevant for the interpretation of our

relatively low PPV (i.e. high number of false positives in relation to the number of true positives) and therefore low screen-in accuracy (PPV * sensitivity). The ‘screen-in accuracy’ is poor according to the criterion (i.e. <0.49), but when taking into account that false positive cases still form a subgroup with a lower social-emotional development, the ‘screen-in accuracy’ may not be as insufficient. Future (longitudinal) research should investigate whether false positive cases on the ASQ-SE-NL develop differently from true negative cases to find out whether – and to what extent – they actually form a group that warrants closer surveillance.

For clinical practice, we advise a monitoring routine in which children are regularly screened. If a child scores positive, we advise practitioners not to draw conclusions on only one ASQ-SE measurement given the relatively high number of false positive results. Instead, we advise to rescreen the child earlier than the regular monitoring moment (e.g. within 3 months instead of 6 months) and incorporate both the clinical impressions of the practitioner and potential concerns of the parent. If the child does not improve, s/he should be referred for further assessment. If the child scores negative on the initial screening, the regular monitoring routine can be followed.

The current study contains several strengths. The participants in our sample were equally distributed over the different ASQ-SE-NL age versions and covered all Bayley-III-NL-SE age versions. Furthermore, it concerned a community sample consisting of mostly healthy children. This fits the population for which the ASQ-SE-NL is designed. This study has some limitations. Firstly, both instruments were filled out by the same informant, which might have led to an overestimation of the agreement between the instruments. On the other hand, in practice both instruments are usually also answered by one informant making the results generalizable to practice. Secondly, as Dutch norms for the ASQ-SE have not been established before, we created cut-off scores based on our sample, but these are different for each sample.

In conclusion, the ASQ-SE-NL and Bayley-III-NL-SE often identify the same children as having *no* delay or problems. Using the Bayley-III-NL-

Table 5

Results of comparisons between ASQ-SE-NL and Bayley-III-NL-SE for all ages together.

Approach	ASQ-SE-NL cut-off	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	UI+ (%)	UI– (%)	Risk of delay or problems on ASQ-SE-NL (n (% of total))	Delay or problems Bayley-III-NL-SE n (% of total)	(Risk of) delay or problems on both	Total n
1	>1 SD	32.5%	88.5%	34.7%	87.5%	0.11	0.77	150 (14.8%)	160 (15.8%)	52 (5.1%)	1014
2	ROC ^a	70.6%	70.8%	32.7%	92.3%	0.23	0.65	220 (36.2%)	102 (16.8%)	72 (11.8%)	608 ^a

Note: ASQ-SE-NL = the Dutch translation of the Ages and Stages Questionnaire – Social-Emotional; Bayley-III-NL-SE = the Dutch translation of the Social-Emotional Scale of the Bayley-III; PPV = positive predictive value; NPV = negative predictive value; UI+ = “screen-in accuracy”; UI– = “screen-out accuracy”; SD = standard deviation; ROC = Receiver Operating Characteristic.

^a ASQ-SE-NL age versions 6–12 months were not included in this calculations as cut-off scores could not be determined due to the low AUC.

SE as the criterion shows that the ASQ-SE-NL performs well in detecting delay or problems (i.e. sensitivity) in the ASQ-SE-NL ≥ 18 months age versions when using the cut-off scores based on ROC curves, but not when using >1 SD. The clinical utility for screening out children without delays or problems was high, but the 'screen-in accuracy' was low. However, this is partly explained by the fact that we screened a healthy population which naturally results in a higher number of false positives for every true positive and therefore lower scores on the PPV and the screen-in accuracy. In practice, the ASQ-SE-NL can form part of a monitoring routine for toddlers. We advise to rescreen the child earlier than the regular monitoring moment in case of a positive test result. The Bayley-III-NL-SE could be used in a next step to (dis-)confirm the problems found with the ASQ-SE-NL.

Funding

This research was supported by ZonMw project number 15071.3001.

CRedit authorship contribution statement

Lisa J.G. Krijnen: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Marjolein Verhoeven:** Conceptualization, Writing – review & editing, Supervision. **Anneloes L. van Baar:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

None declared.

Acknowledgements

The authors wish to thank all children and parents that participated in this study.

References

- J.M. Love, E.E. Kisker, C. Ross, H. Raikes, J. Constantine, K. Boller, J. Brooks-Gunn, R. Chazan-Cohen, L.B. Tarullo, C. Brady-Smith, A.S. Fuligni, P.Z. Schochet, D. Paulsell, C. Vogel, The effectiveness of early head start for 3-year-old children and their parents: lessons for policy and programs, *Dev. Psychol.* 41 (2005), <https://doi.org/10.1037/0012-1649.41.6.88>.
- A.J. Reynolds, J.A. Temple, S.R. Ou, D.L. Robertson, J.P. Mersky, J.W. Topitzes, M. D. Niles, 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.* 161 (2007), <https://doi.org/10.1001/archpedi.161.8.730>.
- P.H. Lipkin, M.M. Macias, Promoting optimal development: identifying infants and young children with developmental disorders through developmental surveillance and screening, *Pediatrics*. 145 (2020), <https://doi.org/10.1542/peds.2019-3449>.
- J.H. Gilmore, R.C. Knickmeyer, W. Gao, Imaging structural and functional brain development in early childhood, *Nat. Publ. Gr.* 19 (2018), <https://doi.org/10.1038/nrn.2018.1>.
- N. Eisenberg, Emotion, regulation, and moral development, *Annu. Rev. Psychol.* 51 (2000), <https://doi.org/10.1146/annurev.psych.51.1.665>.
- C. Saarni, The development of emotional competence, in: *Hanb. Emot. Intell. Theory, Dev. Assessment, Appl. Home, Sch. Work*, 2000.
- T. Yates, M.M. Ostrosky, G.A. Cheatham, A. Fetting, L. Shaffer, R.M. Santos, Research synthesis on screening and assessing social-emotional competence, in: *Center on the Social Emotional Foundations for Early Learning*, 2008 [cited 2021 April 19]. Available from, http://csefel.vanderbilt.edu/documents/rs_screening_g_assessment.pdf.
- F.P. Glascoe, Parents' evaluation of developmental status: how well do parents' concerns identify children with behavioral and emotional problems? *Clin. Pediatr. (Phila)*. 42 (2003) <https://doi.org/10.1177/000992280304200206>.
- H. Hix-Small, K. Marks, J. Squires, R. Nickel, Impact of implementing developmental screening at 12 and 24 months in a pediatric practice, *Pediatrics*. 120 (2007), <https://doi.org/10.1542/peds.2006-3583>.
- J. Squires, D. Bricker, E. Twombly, K. Murphy, R. Hoselton, M.S. Brookes, ASQ:SE-2 technical report excerpted from ASQ:SE-2™ User's guide. www.brookespublishing.com, 2015.
- S.J. Meisels, S. Provence, Screening and assessment: Guidelines for identifying young disabled and developmentally vulnerable children and their families, in: *National Center for Clinical Infant Programs*, 1989, pp. 13–14.
- J. Squires, D. Bricker, E. Twombly, The Ages and Stages Questionnaires: Social-Emotional (ASQ: SE) User's Guide, Paul H, Baltimore, 2002.
- K.H. Armstrong, H.C. Agazzi, The Bayley-III Cognitive Scale, in: *Bayley-III Clin. Use Interpret*, Elsevier Inc., 2010, pp. 29–45, <https://doi.org/10.1016/B978-0-12-374177-6.10002-9>.
- A. Yue, Q. Jiang, B. Wang, C. Abbey, A. Medina, Y. Shi, S. Rozelle, Concurrent validity of the ages and stages questionnaire and the Bayley scales of infant development III in China, *PLoS One* 14 (2019), <https://doi.org/10.1371/journal.pone.0221675>.
- N. Bayley, *Bayley Scales of Infant and Toddler Development: Administration Manual*, 3rd ed., Harcourt Assessment, San Antonio, Tex, 2006.
- M.S. De Wolff, M.H.C. Theunissen, A.G.C. Vogels, S.A. Reijneveld, Three questionnaires to detect psychosocial problems in toddlers: a comparison of the BITSEA, ASQ:SE, and KIPPI, *Acad. Pediatr.* 13 (2013) 587–592, <https://doi.org/10.1016/j.acap.2013.07.007>.
- M.H.C. Theunissen, A.G.C. Vogels, M.S. De Wolff, M.R. Crone, S.A. Reijneveld, Comparing three short questionnaires to detect psychosocial problems among 3 to 4-year olds, *BMC Pediatr.* 15 (2015) 1–8, <https://doi.org/10.1186/s12887-015-0391-y>.
- T.M. Achenbach, C. Edelbrock, *Manual for the Child Behavior Checklist (4–18) and 1991 Profile*, University of Vermont, Department of Psychiatry, Burlington, VT, 1991, p. 1991, in: Google Sch., Lawrence Erlbaum Associates Publishers, 1991.
- A.L. Van Baar, L.J.P. Steenis, M. Verhoeven, D.J. Hessen, Bayley-III-NL. Technische Handleiding, Pearson Assessment and Information, B.V., Amsterdam, The Netherlands, 2014.
- F.P. Glascoe, Are overreferrals on developmental screening tests really a problem? *Arch. Pediatr. Adolesc. Med.* 155 (2001) 54–59, <https://doi.org/10.1001/archpedi.155.1.54>.
- F.P. Glascoe, Early detection of developmental and behavioral problems, *Pediatr. Rev.* 21 (2000), <https://doi.org/10.1542/PIR.21.8-272>.
- L.J.P. Steenis, M. Verhoeven, D.J. Hessen, A.L. van Baar, Parental and professional assessment of early child development: the ASQ-3 and the Bayley-III-NL, *Early Hum. Dev.* 91 (2015) 217–225, <https://doi.org/10.1016/j.earlhumdev.2015.01.008>.
- L.J.P. Steenis, M. Verhoeven, D.J. Hessen, A.L. van Baar, First steps in developing the Dutch version of the Bayley III: is the original Bayley III and its item sequence also adequate for Dutch children? *Eur. J. Dev. Psychol.* 11 (2014) 494–511, <https://doi.org/10.1080/17405629.2013.869207>.
- 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 (2015), e0132871, <https://doi.org/10.1371/journal.pone.0132871>.
- Centraal Bureau voor de Statistiek, *Standaard Onderwijsindeling 2021*, (n.d.). <https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/onderwijs-en-beroe-pen/standaard-onderwijsindeling-soi-/standaard-onderwijsindeling-2021> (accessed April 23, 2021).
- C. Breinbauer, T.L. Mancil, S. Greenspan, The Bayley-III Social-Emotional scale, in: *Bayley-III Clin. Use Interpret*, Elsevier Inc., 2010, pp. 147–175, <https://doi.org/10.1016/B978-0-12-374177-6.10005-4>.
- N. Bayley, G.P. Aylward, *Bayley Scales of Infant and Toddler Development. Technical Manual*, 4th ed., Psychorp., United States of America, 2019.
- A. Evers, W. Lucassen, R. Meijer, K.S. Nip, COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie), Amsterdam. <https://dare.uva.nl/document/2/79346>, 2009 (accessed April 23, 2021).
- IBM, *IBM SPSS Statistics Software for Windows, Version 25*, IBM, 2017.
- E. Colak, F. Mutlu, C. Bal, S. Oner, K. Ozdamar, B. Gok, Y. Cavusoglu, Comparison of semiparametric, parametric, and nonparametric ROC analysis for continuous diagnostic tests using a simulation study and acute coronary syndrome data, *Comput. Math. Methods Med.* 2012 (2012), <https://doi.org/10.1155/2012/698320>.
- W.J. Youden, Index for rating diagnostic tests, *Cancer*. 3 (1950) 32–35, [https://doi.org/10.1002/1097-0142\(1950\)3:1<32::AID-CNCR2820030106>3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3).
- J.E. Fischer, L.M. Bachmann, R. Jaeschke, A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis, *Intensive Care Med.* 29 (2003) 1043–1051, <https://doi.org/10.1007/s00134-003-1761-8>.
- M.J. Borst, Evidence-based practice: the basic tools, in: *Cooper's Fundam. Hand Ther*, Mosby, 2020, pp. 15–20, <https://doi.org/10.1016/B978-0-323-52479-7.00002-8>.
- S. Tenny, M.R. Hoffman, Prevalence, StatPearls Publishing, 2020. Treasure Island (FL), University of Nebraska Medical Center, <http://europepmc.org/abstract/MED/28613617>.
- A.J. Mitchell, How to: analyse a screening or diagnostic study, n.d. www.psychoncology.info (accessed April 23, 2021).
- A.J. Mitchell, The 3 item anxiety subscale of the Edinburgh postpartum depression scale may detect postnatal depression as well as the 10 item full scale, *Evid. Based. Ment. Health*. 12 (2009) 44, <https://doi.org/10.1136/ebmh.12.2.44>.
- A.J. Mitchell, The clinical significance of subjective memory complaints in the diagnosis of mild cognitive impairment and dementia: a meta-analysis, *Int. J. Geriatr. Psychiatry* 23 (2008) 1191–1202, <https://doi.org/10.1002/gps.2053>.
- A.J. Mitchell, Sensitivity \times PPV is a recognized test called the clinical utility index (CUI+), *Eur. J. Epidemiol.* 26 (2011) 251–252, <https://doi.org/10.1007/s10654-011-9561-x>.
- J. Crowell, Development of emotion regulation in typically developing children, *Child Adolesc. Psychiatric Clin. North Am.* 30 (2021). <https://europepmc.org/article/med/34053680> (accessed July 1, 2021).
- G. Dixon, N. Badawi, D. French, J.J. Kurinczuk, Can parents accurately screen children at risk of developmental delay? *J. Paediatr. Child Health* 45 (2009) 268–273, <https://doi.org/10.1111/j.1440-1754.2009.01492.x>.