University of Groningen

Low verbal assessment with the Bayley-III

Visser, Linda; Ruiter, Selma; van der Meulen, Bieuwe; Ruijsenaars, Wied; Timmerman, Marieke

Published in:
Research in Developmental Disabilities

DOI:
[10.1016/j.ridd.2014.09.014](https://doi.org/10.1016/j.ridd.2014.09.014)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2015

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Visser, L., Ruiter, S., van der Meulen, B., Ruijsenaars, W., & Timmerman, M. (2015). Low verbal assessment with the Bayley-III. *Research in Developmental Disabilities*, 36, 230-243.
<https://doi.org/10.1016/j.ridd.2014.09.014>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Low verbal assessment with the Bayley-III

Linda Visser^{a,*}, Selma A.J. Ruiter^b, Bieuwe F. Van der Meulen^a,
Wied A.J.J.M. Ruijsenaars^a, Marieke E. Timmerman^c

^a University of Groningen, Faculty of Behavioural and Social Sciences, Department of Special Needs Education and Youth Care, Grote Rozenstraat 38, 9712TJ Groningen, The Netherlands

^b De Kinderacademie Groningen, Herestraat 106, 9711LM Groningen, The Netherlands

^c University of Groningen, Faculty of Behavioural and Social Sciences, Department of Psychometrics and Statistics, Grote Kruisstraat 2/1, 9712TS Groningen, The Netherlands

ARTICLE INFO

Article history:

Received 3 June 2014

Received in revised form 17 September 2014

Accepted 22 September 2014

Available online

Keywords:

Developmental assessment

Young children

Language impairment

Accommodations

ABSTRACT

Recently, the authors have developed the Bayley-III-NL Low Verbal for developmental assessment of children with language impairment. The Low Verbal version consists of an accommodated cognition scale, and non-accommodated communication and motor scales. The purpose of the study was to evaluate the validity and added value of the Bayley-III-NL Low Verbal for children with a language impairment, in relation to the standard Bayley-III-NL for children without impairment. We administered the Bayley-III Low Verbal to 69 children with language impairment, and the standard Bayley-III-NL to 1132 children without impairments. We used an evaluation form for test administrators and interviews with developmental psychologists to evaluate the suitability of the Low Verbal version for the target group. We analyzed the test results using nonparametric item response theory (IRT) to investigate whether test results can be reasonably compared across the two groups. The results of the IRT analyses support the validity of the Bayley-III-NL Low Verbal: the test items do not suffer from differential item functioning (DIF) across the two groups, and thus measure the ability levels of interest in the same way. The results of the evaluation form and interviews confirm that the Bayley-III-NL Low Verbal has added value for testing children with a language impairment, especially for children up to 36 months old. It is also suitable for children with general developmental delay. We conclude that the Bayley-III-NL Low Verbal can validly assess the cognitive, language, and motor development of young children with a language impairment and is the preferred instrument for this target group.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Many children receiving early intervention services have a language impairment. Mashburn and Myers (2010), for example, report that this prevalence is 62%. A language impairment is defined as a disorder in one or more of the basic learning processes involved in understanding or in using spoken or written language (FL State Board of Education, 2010). As a consequence of the interrelatedness between language and cognitive development (Thal, 1991), many children with

* Corresponding author. Tel.: +31 503636556.

E-mail addresses: Linda.Visser@rug.nl (L. Visser), Selma.Ruiter@dekinderacademie.com (Selma A.J. Ruiter), B.F.van.der.Meulen@rug.nl (B.F. Van der Meulen), A.J.J.M.Ruijsenaars@rug.nl (Wied A.J.J.M. Ruijsenaars), M.E.Timmerman@rug.nl (M.E. Timmerman).

language impairment also have certain cognitive impairments, especially in cases of receptive language problems (Botting, 2005; Elbro, Dalby, & Maarbjerg, 2011; Van Daal, Verhoeven, & Van Balkom, 2009). Children with language impairment form a very heterogeneous group because they differ widely with respect to levels of receptive and expressive language, as well as cognitive development (Dodd & Crosbie, 2002).

Language impairment is associated with worse outcomes in the longer term in domains such as communication, cognition, educational attainment, employment, and socioeconomic status (Elbro et al., 2011; Johnson, Beitchman, & Brownlie, 2010). This is especially so when a child has both receptive and expressive communication problems (McCabe, 2005) or additional problems in areas other than language (Elbro et al., 2011). Early intervention can play an important role in reducing the potential negative consequences of early language impairment (Guralnick & Conlon, 2007). Early intervention generally involves developmental assessment, which enables optimal tuning of the subsequent intervention to the child.

When applying developmental assessment, the heterogeneity of the group of children with language impairment should be taken into account. This means that domains other than language should be considered in the assessment because of the high risk for comorbid problems. Also, it means that the instruments used should be valid in cases of language impairment (Crais, 2009). The risk is, however, that a language impairment influences test results on measures of development on other domains, like cognition and motor skills. This would result in differential item functioning (DIF): the situation in which differences exist in the way a test item functions across certain groups that are matched on the attribute measured by that item (Osterlind & Everson, 2009). In other words, an item that is meant to measure cognition, could give a different score for a child with language impairment than for a child without language impairment, even though these two children have the same level of cognitive development. DIF threatens the construct validity of an instrument, that is, the extent to which the instrument actually measures the intended construct.

A widely used instrument for developmental assessment in early intervention is the *Bayley Scales of Infant and Toddler Development* (3rd ed.) (Bayley-III; Bayley, 2006a). The current study focuses on the suitability of the Dutch version of the Bayley-III (Bayley-III-NL; Steenis, Verhoeven, & Van Baar, 2012) for assessing children with a language impairment. The Bayley-III is the most recent successor of the BSID-II (Bayley, 1993; Van der Meulen, Ruiter, Spelberg, & Smrkovsky, 2002) and has improved test materials and an adjusted structure for the developmental scales. The psychometric qualities appear to have been improved over its previous version and updated norms are available (Bayley, 2006b). It is a widely used and examined instrument for developmental assessment of children between 0.1 and 3.6 years of chronological age. The Bayley-III is also being used with older children who have an expected developmental age of less than 3.6 years. The instrument assesses development in the areas of cognition, receptive communication, expressive communication, fine motor development, and gross motor development.

The cognition scale appeared to contain many language components: for many items, questions are asked to the child or instructions given, which means that the child needs a good language comprehension to understand the task at hand and be able to show his or her cognitive abilities. Therefore, we developed a Low Verbal version of the Bayley-III-NL. The term “Low” refers to the number of language components in the test items. We accommodated the cognition scale: our objective was to lower the degree of verbal ability needed to perform the test items, while keeping the item content and difficulty unchanged (Thurlow, Elliot, & Ysseldyke, 2003). This would then imply a measure of cognitive development – i.e. the development of a set of processes in order to gain knowledge of the world, by means of thinking, learning, perceiving, remembering and understanding – that is minimally dependent on language abilities and hence has a higher construct validity for children with a language impairment. We did not accommodate the Communication scales of the Bayley-III-NL, because these assess the language skills of a child. Lowering the number of language components in the items would therefore change the content of the scale and thus threaten the construct validity. We also did not accommodate the motor scales, because the items belonging to the motor scales rely to a very limited extent on language. This is caused by the fact that items from the motor scales are mainly based on observation (e.g., “Lifts cup by handle”) or imitation (e.g., “Balances on left foot”). If research shows that the item content and difficulty have indeed remained the same in the Low Verbal version compared to the standard version for all scales, the standard norms would remain applicable.

This paper describes the results of a validity study on the Bayley-III-NL Low Verbal. We used item response analysis to investigate if (1) the Bayley-III-NL Low Verbal (consisting of the accommodated cognition scale and non-accommodated communication and motor scales) sufficiently distinguishes among children with different levels of the latent trait, and (2) the Bayley-III-NL Low Verbal contains items that function similarly for children with a language impairment as the corresponding items in the standard Bayley-III-NL for children without a language impairment. Moreover, we used an evaluation form and interviews with developmental psychologists to assess to what extent the Bayley-III-NL Low Verbal is more suitable than the standard version for assessing the development of young children with a language impairment.

2. Materials and methods

The current study consisted of a pilot study and a main study. The pilot study and its results will be described in Section 2.2.2, as a foundation for the Bayley-III-NL Low Verbal version that was used in the main study. For purposes of clarity, the remainder of Section 2 (Materials and Methods), Section 3 (Results) and Section 4 (Discussion), will focus purely on the main study.

2.1. Participants

The main study contained two samples: a special needs group of children with a language impairment and a control group.

Developmental psychologists of 17 different branches of organizations referred children for the special needs group on the basis of the following inclusion criteria: The child has a language impairment, has an expected developmental age between 0.1 and 3.6 years (i.e., the developmental age range of the Bayley-III), and has the minimum abilities required for the test, namely to sit upright in a (wheel)chair and to use at least one hand. Excluded were children with deafness and/or severe visual impairment, so test materials could be well perceived. All referred children were tested and included in the sample of the current study. 25 children were referred by an organization that specifically provides early intervention services to children with a language or hearing impairment; 29 children were in a special day care service for young children with developmental disabilities; 4 children were referred by a primary school for children with intellectual disabilities; 11 children were referred by an organization that provides a wide range of services in the form of parenting support and early intervention. We received information about nonresponse from six of the 17 organizations via evaluative phone calls. Three of these six organizations reported that there were between one and four children who could have been referred, but were not. The reasons reported were difficult communication with parents as a consequence of a language barrier, parents' resistance to testing, health problems of the child, and the child moving to a different organization.

The sample consisted of 69 children, of which 26 girls and 43 boys. The mean age was 4.8 years, with a standard deviation of 1.11 years. The range was rather wide: 2.2–10.8 years. The developmental age of the children according to their score on the cognition scale of the Bayley-III using the United States standardization was 2.4 years, with a standard deviation of 0.10 years, a minimum of 0.4 and a maximum of 3.6 years.

The developmental psychologist from the referring organization provided information about the presence of general developmental delay and the diagnoses and impairments of the child. Among the 69 children, 59 had a general developmental delay, eight had an expected average cognitive development, and for two children we did not have information about the developmental level in advance. All children in this sample had a language impairment. We have summarized the information concerning diagnoses and impairments of the children in Table 1.

The control group consisted of 1132 children from different parts of the Netherlands. The children all had an expected development within the normal range, no known impairments, and a chronological age between 0.1 and 3.6 years. The average age was 1.7 years, with a standard deviation of 1.0 years, and a range of 0.1–3.6 years. There were 547 girls and 585 boys. The control group consisted of all children who were included in the sample of the Bayley-III-NL standardization research up to April 11, 2013. The standardization study ran from 2010 to mid 2014.

The chronological age of the children in the special needs group was higher than that of the children in the control group. This is a consequence of the fact that many children with a language impairment also have a cognitive impairment and therefore their chronological age is higher than their developmental age. The special needs group is representative of children in daily practice who are assessed with the Bayley-III-NL in organizations supporting children with a language impairment (amongst others). The control group is representative for the population of Dutch children in the age range of 0–3.6 years.

2.2. Instruments

The children from the control group were tested with the standard Bayley-III-NL; the children from the special needs group were tested with the Bayley-III-NL Low Verbal, which consists of the accommodated cognition scale and the non-accommodated communication and motor scales.

Table 1

Diagnoses and impairments: special needs group, main study.

Diagnosis ^a	Impairment				Total
	LI	LI & motor	LI & visual & motor	Unknown	
None	29	3	0	0	32
PDD	9	4	0	0	13
Down syndrome	5	2	1	0	8
ADHD	2	0	0	0	2
Angelman syndrome	0	1	0	0	1
Other genetic disorder	2	0	1	0	3
Other	0	4	1	0	5
Hydrocephalus & other	0	1	0	0	1
PDD & other	1	0	0	0	1
Other genetic disorder & other	0	1	0	0	1
Unknown	1	0	0	1	2
Total	49	16	3	1	69

Note: LI – language impairment; PDD – pervasive developmental disorder; ADHD – attention deficit hyperactivity disorder; Other includes Dandy–Walker syndrome, tuberous sclerosis, diffuse myelination disorder, fragile X syndrome, Arnold–Chiari malformation, 17q11.2 deletion, and epilepsy.

^a The numbers include children who had already received a diagnosis as well as children who were still in the assessment process.

2.2.1. Bayley-III-NL

The Bayley-III (Bayley, 2006a) is an individually administered instrument for developmental assessment of children in the chronological age range of 0.1–3.6 years. It is also being used with older children who have an expected developmental age within this range. The instrument assesses the development in the areas of cognition (91 items), receptive communication (49 items), expressive communication (48 items), fine motor development (66 items), and gross motor development (72 items). The scales are built up of items increasing in difficulty. Items are scored dichotomously: a child receives a score of 1 when he/she passes an item and 0 when he/she fails an item. The starting point depends on the chronological age of the child. Items before the starting point are not administered and automatically scored 1. Regarding the basal and ceiling rules, the basal rule is that the first three administered items should have a score of 1. If not, the test administrator should go back to the previous starting point until the rule has been met. The ceiling rule is to stop after five consecutive items have been scored 0; all items after the final administered item are not administered and automatically scored 0. A raw score can be calculated for each scale, counting all items that scored 1. For children with a chronological age of up to 3.6 years, a standard score (average 10, SD 3) and index score (average 100, SD 15) can be calculated on the basis of this raw score. For children with a chronological age of more than 3.6 years, a developmental age equivalent can be looked up for each scale, which corresponds to the raw score. The domains of social–emotional development and adaptive behavior can be assessed on the basis of primary caregiver responses to a questionnaire, which was not part of this current research.

The standardization sample of the Bayley-III in the United States included 1700 children. Validity data are given in the form of moderate to high correlations of Bayley-III test scores with scores on other instruments. The internal consistency coefficients range from 0.71 to 0.98 across age groups and scales, with an average of 0.89. Test–retest stability coefficients range from 0.67 to 0.94 across age groups and scales and increase with age (Bayley, 2006b).

The Bayley-III-NL is different from the United States version in terms of language used (Dutch rather than English) and two deleted items in the expressive communication scale, resulting in a total of 46 items. The two items were deleted, because they appeared to be unsuitable with respect to Dutch language development.

2.2.2. Bayley-III-NL Low Verbal: characteristics and development

The Bayley-III-NL Low Verbal has the same characteristics as the standard version of the instrument, except for the accommodations. The manual and the record form differ from the standard Bayley-III-NL, whereas the test materials are the same. The accommodations to the cognition scale are based on those of the BSID-II-NL Nonverbal version (Ruiter, Hoekstra, Van der Meulen, Lutje Spelberg, & Nakken, 2006) and can be divided into a procedural accommodation and item accommodations. The procedural accommodation means that the test administrator should administer the items with emphasis: using short sentences, talking clearly, and placing extra emphasis on the key words when talking to the child. This is described in the Low Verbal manual. The item accommodations include the addition of commonly used signs to items in which language comprehension of the child is necessary in the standard version. An example is the thumbs-up gesture to indicate “Well done!”. An instruction was also added to applicable items describing how to attract the attention of the child. These signs and instructions are described explicitly in the Low Verbal manual per item, as an addition to the item instructions of the standard version of the instrument, providing an optimal degree of standardization. The record form contains a pictogram before each item that contains a Low Verbal accommodation so that the test administrator is reminded to take the accommodations into account. When a commonly used sign is applicable, the record form provides a pictogram or short word referring to that particular sign so that the test administrator does not need the manual during testing. See Fig. 1 for an example of a few items in the record form.

2.2.2.1. Pilot results. We named the first version of the accommodated instrument that we developed the “Nonverbal” version. We performed a pilot study in which we assessed the suitability of the Nonverbal version for children in the target group. The sample consisted of 11 children (who are not included in the main sample) with a language or hearing impairment and a mean age of 3.4 years (SD 1.1 years; range 1.8–5.8 years). The children in the pilot study were tested with both the Bayley-III-NL Nonverbal and the standard Bayley-III-NL, in randomly assigned counterbalanced order, so that the test administrator could compare the two testing sessions. Two children were only tested with the Nonverbal version. The test administrators filled in an evaluation form about the suitability and practicality of the accommodations after each test administration with the Nonverbal version. In addition, semi-structured interviews were performed with two experts who worked as developmental psychologists with children with a language or hearing impairment.

The children obtained an average raw score on the standard cognition scale of 63.4 (SD 10.4; range 45–76) and on the Nonverbal cognition scale 65.3 (SD 9.5, range 50–79). Five of the children obtained 3–5 points higher on the Nonverbal version than on the standard version; for the other children the score difference was 0 or 1.

The results of the evaluation form showed that the accommodations were suitable and practical in the majority of cases (10 out of 11 cases). For three children, the suitability was reported not to be specifically caused by the accommodations. These children had limited communication abilities due to pervasive developmental disorder and, as a consequence, did not have sufficient communication abilities to acknowledge the gestures. The seven children for whom the accommodations were reported to be beneficial obtained an equal or higher score on the accommodated version as compared to the standard version (0–3 raw score points higher), in contrast to the child for whom the Bayley-III-NL Nonverbal was not suitable, who scored 1 raw score point lower. This child was deaf and the accommodations were not sufficient to compensate for this.

	ITEM	MATERIALS	SCORE CRITERIA AND COMMENTS	SCORE	GESTURE
36 54	54. Block Series: 9 blocks	9 blocks Cup with handle	Trials:1 Child places all 9 blocks inside cup at one time. # of blocks in cup:	1 0	In ?
47 55	55. Pegboard series: 6 pegs	Peg board 6 yellow pegs	Trials: 3 Child places all 6 pegs in pegboard within 70 seconds. Trial 1 Completion time: Trial 2 Completion time: Trial 3 Completion time:	1 0	> < ?
49 56	56. Pink board series: Completes	Pink board Red block set (square, circle, triangle)	Trials:1 Child correctly places all three pieces within 180 seconds. # pieces placed correctly (180 seconds):	1 0	?
	57. Used pencil to obtain object	Pencil Small red duck	Trials: 2 Child uses pencil to attempt to obtain duck.	1 0	Nod

Fig. 1. Example of items in the Low Verbal record form. Items 54–57 in the research version of the Bayley-III-NL Low Verbal record form. Explanation of gestures as written in the manual, from top to bottom: 'In gesture' – make a round opening with the index finger and thumb of the left hand and point with the index finger of the other hand into the opening; 'Questioning gesture' – open hands with hand palms pointing upwards, put hands in front of body. Move hands in the direction of the child, nodding to the child at the same time; 'Good gesture' – thumb up (can always be applied when applicable); 'Open hand gesture' – to point at materials, use an open hand with palm of the hand directed upwards, point to the materials with the index finger and move hand along the materials; 'Together gesture' – first point at both materials individually, then move both hands toward each other with the palm of the hands pointed downwards; 'Nod gesture': stretch arm forwards with the palm of the hand upwards and bend your fingers.

The outcome of the two interviews confirmed this last point. The two experts indicated that the Bayley-III-NL Nonverbal is not suitable for children above 3 years of age who are deaf, because the test items higher on in the scale still contained too much language and no instructions for using sign language are available. They also indicated that the Nonverbal cognition scale still contained too much language to be called a nonverbal test. This was caused by the fact that we did not remove any items from the test. This was to ensure that the standard norms remained applicable. The experts did find the accommodated version suitable for children with a language problem, however.

During the pilot study, it appeared that administering the standard Bayley-III-NL to a child with a language impairment was not useful. The test administrators regularly indicated that a very awkward situation arose when they tried to administer the standard Bayley-III-NL to a child from the target group, because it came across as very unnatural and insensible when they restrained from using supportive gestures while they were actually needed. They also reported occasionally and unintentionally incorporating (unstandardized) adjustments in the standard test administration, such as the use of gestures.

On the basis of the results from the pilot study, we adjusted the accommodated instrument and planned the main research. We added information to the record form and manual. For example, we added pictures of the gestures and a note that the items in the higher age range (above 3 years) still require a verbal answer from the child. We renamed the instrument "Low Verbal." For the main study, we decided to administer only the Low Verbal version and not the standard version to the children. Furthermore, we decided to concentrate the main study on the suitability of the Bayley-III-NL Low Verbal for children with a language impairment (possibly combined with a hearing impairment) because the instrument appeared to be especially suitable for children in this group and not for children with deafness.

2.2.3. Instruments for qualitative evaluation of the Bayley-III-NL Low Verbal

The test administrators in the main study filled in an evaluation form for each child in the special needs group. They were asked about the suitability of the Bayley-III-NL Low Verbal for the specific child and about possible errors in the material. They used their experience with the standard version of the Bayley-III-NL as well as the Low Verbal version to judge to what extent the Low Verbal version was suitable. The test administrators incorporated the opinion of the referring psychologist, who knew the child well, in their evaluation.

To obtain an overview of the opinion of developmental psychologists about the suitability and added value of the Bayley-III-NL Low Verbal, a semi-structured interview was done with four developmental psychologists who worked with young children with a language impairment. These interviews were conducted after the tests for the current research were completed. All the psychologists had already had experience with the Bayley-III-NL Low Verbal by observing one or more test administrations with the instrument on a child they knew, and by interpreting the results of that test administration. One respondent worked at an organization supporting young children with developmental disabilities and their families,

while three of the respondents worked at an organization supporting young children with a language impairment. The respondents had been working as developmental psychologists for between six months and ten years.

The structure of the interviews was based on the following six questions. The interviewers asked additional questions to clarify the answers when necessary.

- (1) To what extent do you think the Bayley-III-NL is suitable for the target group of your organization?
- (2) What is your opinion about the practicality and the added value of the Low Verbal accommodations?
- (3) Does the Bayley-III-NL Low Verbal have added value, compared to the standard version, for the target group of your organization?
- (4) For which children does the Bayley-III-NL Low Verbal specifically have added value, compared to the standard version?
- (5) How did the course of the administration of the standard communication and motor scales of the Bayley-III-NL go, and to what extent did the results give a good picture of the development of the children in these areas?
- (6) What do you find are the advantages and disadvantages of the Bayley-III-NL Low Verbal when compared to the Snijders-Oomen Nonverbal Intelligence Test – Revised (SON-R 2.5–7; [Tellegen, Winkel, Wijnberg-Williams, & Laros, 1998](#))? The SON-R 2.5–7 is a widely used instrument in the Netherlands for assessing nonverbal intelligence.

2.3. Procedure

2.3.1. Special needs group

We received approval for the pilot and main research from the Medical Ethical Committee of the University Medical Center Groningen, located in the Netherlands.

The referring developmental psychologist filled in a referral form with information about the child, such as expected developmental level, diagnoses, and impairments. Permission was asked from the parents via a consent form. The child was then tested at the organization, which was a familiar location, because the child attended it multiple days a week. Two children were tested by the referring psychologist; the other children were tested by a test administrator from the university. These test administrators were advanced university students in special needs education or psychology. All test administrators had received training in administering the standard and Low Verbal versions of the Bayley-III-NL. The training program included one 4-h session, during which the participants received information about the instrument and completed a case example of a test administration. The students also performed five test administrations with children from their own circle of acquaintances in order to practice. They recorded two of these five test administrations and received feedback from one of the researchers, who was very experienced with the Bayley-III-NL.

After the test administration, the test administrator filled in the evaluation form and wrote a test report. The report and test results were given to the referring psychologist so that they could be used in the regular developmental evaluation of the child.

Four semi-structured interviews were performed with developmental psychologists who had been involved in the research. Two of those interviews were administered by one of the researchers, and two by two different test administrators. The interview questions were answered in written form by the respondent before the actual interview, and the interviewer then complemented these written answers during and immediately after the interview with information obtained by asking additional questions.

2.3.2. Control group

The tests of the children in the control group were administered as part of the standardization research being done in the Netherlands, which has been performed by colleagues from Utrecht University. Separate approval was obtained for this research via the Medical Ethical Committee of the University Medical Center Utrecht. Parents of children who met the inclusion criteria were approached using flyers and information letters via multiple channels, including infant welfare centers, child day-care centers, and local governments across the entire Netherlands. Parents of children who agreed to participate in the research by returning a consent form were invited for a test administration. Some children were tested at home ($n = 176$), but most children were tested at their day-care center or a different location outside their home, such as a testing room in an arranged location near the child's home.

2.3.3. Exclusion of non-reliable results

The aim was to administer all five scales of the Bayley-III-NL to all children in the special needs group as well as in the control group. This did not always happen, with two causes. First, the Bayley-III is an extensive test with an average 90-min duration when given to older children. Consequently, it appeared not always possible to administer all the scales within the time that was available. When the information about language or motor skills was not necessary for the support of the child in clinical practice, no second assessment was planned in case some of the scales were not administered during the initial assessment meeting. Second, we categorized all the scales with more than 3 not-scored items (e.g., refused by the child or forgotten by the test administrator) as non-reliable. This resulted in the following numbers of reliable test administrations per scale. For the special needs group, 67 children were reliably administered the cognition scale, 58 the receptive communication, 57 the expressive communication, 60 the fine motor development, and 48 the gross motor development

scale. No children were excluded, because none of the children had a non-reliable administration for all the administered scales. For the control group, 1073 children were reliably administered the cognition scale, 1063 the receptive communication, 1088 the expressive communication, 1078 the fine motor, and 1066 the gross motor development scale. For four children, all the scales appeared to have been administered non-reliably. We excluded these children when calculating the total number of children in the sample ($n = 1132$), and the corresponding average and range of the chronological age.

2.4. Analysis

The analyses for the main study consisted of a quantitative part, associated with the test results, and a qualitative part, associated with the evaluation form and interviews.

2.4.1. Quantitative

2.4.1.1. Descriptive statistics. We started with a descriptive analysis of the test scores of the children in the special needs group and in the control group. We summarized the mean, standard deviation, and the range of the test scores per scale, using the scoring rules of the Bayley-III. Furthermore, we considered the number of times each item was actually administered. It was important to take this into account when interpreting the results, because an item analysis partly based on imputed values – albeit according to the scoring rules of the Bayley-III – might yield biased results. For each item, in each target group, we required an absolute minimum of 20 observed scores in order to draw any conclusion on that specific item in the target group involved.

2.4.1.2. Basic explanation of item response theory (IRT) modeling. To analyze the item scores, we used nonparametric item response theory (IRT). IRT modeling can be used to investigate the properties of a set of items, including underlying traits and the reliability of the resulting scale. The core of any IRT model is the item response function (IRF), which is the success probability for an item as a function of the latent trait score. The properties of the IRFs are used to study the characteristics of items and a set of items that are intended to measure the same latent trait. Specific response probabilities for each specific person–item combination are modeled, on the basis of which conclusions can be drawn about the characteristics of the items (Sijtsma & Molenaar, 2002). In each of the IRT analyses performed, the latent trait refers to the ability aimed to be measures with the scale considered (i.e., cognition, receptive communication, etc.). Note that the chronological and developmental age of the children do not play any explicit role in the IRT analyses; of course, there is a relationship between the latent trait score and age, which could be explored if desired.

In so-called nonparametric IRT models, data do not have to fit a specific parametric form of the IRF. This means that having to force data into a structure that they sometimes do not have can be avoided, resulting in information that is relatively easy to interpret (Meijer & Baneke, 2004). Mokken scale analysis (Mokken, 1971) is a nonparametric item response analysis and our analyses are based on a well-known variant of it: the double monotonicity model (DMM).

2.4.1.3. IRT-analyses in the current study. The Bayley-III is a comprehensive and complex instrument. This complexity increases the chance of data not fitting a parametric model. Therefore, we studied the item characteristics of each of the five scales of the Bayley-III-NL using Mokken scale analysis in MSP5 (Molenaar & Sijtsma, 2000). We made a data file with the dichotomous item scores, applying the scoring rules of the Bayley-III to comply as much as possible with the way the instrument is used in practice. This implies that for items that were not administered to a child due to the basal and ceiling rules, the scores were imputed, where all items before the first administered item were given a score 1 (basal rule), all items after the last administered item were given a score 0 (ceiling rule), and not-scored items (items that were not scored although they should have been scored) were given a score 0. These scoring rules are in line with the fact that the items in the Bayley-III-NL are ordered by difficulty on the basis of the US standardization study (Bayley, 2006a).

We ran a Mokken scale analysis in MSP5 separately for each of the Bayley-III scales cognition, receptive communication, expressive communication, fine motor development, and gross motor development, comparing the special needs group and the control group. We analyzed the output of the Mokken scale analysis in four steps. First, we examined the scalability coefficient of each item (H_i) and the total scale (H). These measures give information about the extent to which items (H_i) and a set of items (H) can discriminate between individuals (i.e., the strength of the items and the scale). They can also be used to assess the assumption of unidimensionality, which means that all items in a scale measure the same latent trait. Following the criterion as set by Mokken (1971), an H_i -value and H -value should be larger than 0.3 to ensure that the item and the scale concerned, respectively, meet the assumption of unidimensionality sufficiently, and have sufficient discriminating power. Because the imputation of scores according to the basal and ceiling rules of the Bayley-III inflates the scalability coefficients, we consider this Mokken criterion of the H -values as the absolute minimum.

Second, we examined the criterion values for monotonicity, as provided by MSP5. The assumption of monotonicity implies that the IRFs are monotonically non-decreasing. In other words: a higher value of the latent trait of a respondent is always related to a higher chance on a positive score on the item. If this assumption does not hold for an item, it means that the item is not suitable to measure the latent trait. A criterion value for monotonicity exceeding 80 is considered a strong indication that an item violates the assumption concerned (Molenaar & Sijtsma, 2000). For each item for which the criterion value appeared to be higher than 80, we inspected the empirical IRF (which is estimated on the basis of the observed data) to examine the source of the non-monotonicity.

Third, we used the criterion value given by MSP5 as a result of the check of non-intersection, using the threshold of 80. The assumption of non-intersection means that it is assumed that the IRF's do not intersect each other. In other words: the order of items in terms of difficulty is the same for respondents of all levels of the latent trait. This assumption is important, because the imputation of item scores according to the scoring rules of the Bayley-III-NL assumes that the items are ordered by increasing difficulty, and that this ordering holds irrespective of the score on the latent trait.

Fourth, we used the criterion values for the equal item-step ordering across groups, again using the threshold of 80. The assumption of equal item-step ordering implies that the item characteristics in terms of their order are the same for the two groups that are being compared (Molenaar & Sijtsma, 2000). The criterion value for this assumption can be used to check if differential item functioning (DIF) exists between the two groups. In the current study, DIF between the special needs group and the control group could originate from structural differences in item functioning between the groups; in the case of the cognition scale, DIF could also stem from differences between the items of the Low Verbal and standard versions, due to the accommodations. If DIF is found, this would indicate that the item has different characteristics when assessed with children with a language impairment and might therefore not be suitable for that target group.

2.4.2. Qualitative

We examined the results on the evaluation forms in detail. We mapped the number of times the test administrators had reported that the child had been able to show his or her abilities, and how often they reported that this was specifically due to the Low Verbal accommodations. This opinion of the test administrators about the suitability of the Low Verbal version for the children in the special needs group is relevant, because it gives information about the added value of the accommodations. This information cannot be extracted from the test results in the current study, which only provided information about the comparability of the test scores to scores of children without an impairment.

To incorporate the view of the developmental psychologists working in the field, we examined the results of the semi-structured interviews. We categorized the different kinds of answers given, mapped the frequencies of the answers, and summarized the most important results.

3. Results

3.1. Test results

3.1.1. Descriptive statistics

We have summarized the descriptive statistics of the test scores of the special needs group in Table 2, and those of the control group in Table 3. The tables show the number of items that each scale contains, and the number of children who were reliably assessed with that scale. The tables also show the average, standard deviation, and the range of the raw scores of the children for each scale. Comparing the control group and the special needs group, we see that there is a larger variation in test scores within the control group, as reflected in the standard deviations. This difference can be explained by the range in chronological age for the special needs group of 2.2–10.8 years: Unlike the control sample, the special needs sample does not include very young children. This is because a language impairment cannot be identified at a very young age where a lack of language comprehension and production is normal. This resulted in relatively few low scores in the special needs sample.

Table 2
Administered items and raw test scores: special needs group.

			Cognition (91 items)	RC (49 items)	EC (46 items)	FM (66 items)	GM (72 items)
Number of times each item was administered	0–9 times	Item nos.	1–33, 37–39, 44, 89–91	1–9, 48–49	1–9	1–25, 63–66	1–43
	10–19 times	Item nos.	34–36, 40–43, 45–59, 87–88	10–14, 16–18, 20–24, 42–47	10–19, 41–46	26–29, 60–62	44–47, 72
	20–29 times	Item nos.	61–62, 82–86	15, 19, 25–27, 38–41	20–26, 28, 36–40	30, 32–34, 53–59	48–53, 67–71
	30–39 times	Item nos.	60, 74–81	30–37	27, 29–35	31, 35–37, 42, 49–52	54–56, 60–66
	≥40 times	Item nos.	63–73	28–29	–	38–41, 43–48	57–59
N (sample)			67	58	57	60	48
Test score			M (SD)	65.6 (13.6)	27.2 (10.0)	27.1 (10.6)	43.2 (9.4)
			Range	26–86	9–45	7–45	22–62
							23–69

Note: RC – receptive communication; EC – expressive communication; FM – fine motor development; GM – gross motor development; item nos. – item numbers; N (sample) – number of children tested with this scale, excluding non-reliable test administrations; M – mean; SD – standard deviation. The upper part of the table shows the number of times the item mentioned has actually been administered to a child in the special needs sample. For items that were not administered, the scores were imputed on the basis of the Bayley-III scoring rules (1 for items before the first administered item and 0 for items after the last administered item).

The lower part of the table shows the number of children from the special needs group that were administered the scale under concern and the corresponding average, standard deviation, and range of the raw test score.

Table 3

Test scores control group.

		Cognition (91 items)	RC (49 items)	EC (46 items)	FM (66 items)	GM (72 items)
<i>N</i> (sample)		1073	1063	1088	1078	1066
Raw score	<i>M</i> (SD)	50.3 (22.9)	21.1 (12.0)	22.4 (14.0)	33.1 (14.7)	42.6 (18.8)
	Range	2–86	2–46	0–46	1–65	3–71
Scale score	<i>M</i> (SD)	10.6 (2.3)	10.0 (2.6)	10.1 (2.4)	11 (2.5)	9.0 (2.9)
	Range	1–19	1–19	2–19	4–19	1–19

Note: RC – receptive communication; EC – expressive communication; FM – fine motor development; GM – gross motor development; *N*(sample) – number of children tested with this scale, excluding non-reliable test administrations; *M* – mean; SD – standard deviation.

The difference between the average raw score of the two groups on the cognition scale was 15.3, with a 95% confidence interval of 11.7–18.9 and a higher average for the special needs group.

To characterize the control group, Table 3 also shows the scale score for each of the five scales of the Bayley-III-NL. The scale score is a standardized score with an average of 10 and a standard deviation of 3 and is based on the standardization sample from the United States, because the Dutch norms are not yet available. The average cognition and Fine motor scale scores appear to be a bit higher than the average of 10, while the average Gross motor scale score is a bit lower than the 10. This could be due to differences in norms between the Netherlands and the United States (Steenis et al., 2012) and/or to the fact that the sample is not yet completely representative of the Dutch population.

Table 2 also shows the number of times each item was administered to one of the children in the special needs group. The following items met the criterion of a minimum of 20 administrations: items 60–86 (cognition scale), items 25–41 (receptive communication scale), items 20–40 (expressive communication scale), items 30–59 (fine motor development scale), and items 48–71 (gross motor development scale). In the control group, all items in all scales met the minimum of 20 administrations.

3.1.2. IRT: main relevant findings

Table 4 summarizes the results of the Mokken scale analysis for the special needs group and the control group, separately. We will now discuss the results following the same four steps as those described in Section 2.4.1.3.

First, the H_i -value of the items per scale and the H -value of each scale of the Bayley-III were all larger than 0.3, indicating that the assumption of unidimensionality has been met sufficiently and the items and scales have sufficient discriminating power.

Second, the criterion values of the check of monotonicity are reported in Table 4 in cases where they were 80 or higher, hinting at a violation of this assumption. The numbers between brackets are the criterion values that were given by MSP in cases of such a violation. Inspection of the IRF's of the items with a violation revealed that they were based on the test results of relatively few children ($n = 16$ –27, depending on the item). In Section 3.1.3, we elaborate on the results of the inspection of IRF's.

Table 4

Results of Mokken scale analysis per group: scalability coefficients and items with violations.

		Cognition	RC	EC	FM	GM
Special needs group	H_i of items: range	0.51–1.00	0.50–0.97	0.77–0.99	0.63–1.00	0.52–0.97
	H of scale	0.84	0.86	0.91	0.83	0.79
	Violation of monotonicity: item (criterion value)	–	19 (109)	–	–	66 (96)
	Violation of non-intersection	–	–	–	–	–
Control group	H_i of items: range	0.79–0.99	0.69–0.97	0.88–0.99	0.77–0.99	0.82–0.99
	H of scale	0.94	0.94	0.97	0.94	0.97
	Violation of monotonicity: items (criterion value(s))	1, 2, 5, 46 (83–141)	49 (87)	–	1 (129)	1 (195)
	Violation of non-intersection: items (criterion value(s))	41,45,46, 50, 53–58, 60, 69 (80–127)	–	–	46 (85)	24 (96)
DIF	Range of criterion values	0–36	0–19	0	0	0–55

Note: RC – receptive communication; EC – expressive communication; FM – fine motor development; GM – gross motor development.

The table shows the (range of) scalability coefficients for the items (H_i) and the scale (H). Also, it shows the items for which a violation of the assumption of monotonicity was found (i.e., the item response function appeared not to be monotonically non-decreasing) and the items for which a violation of the assumption of non-intersection was found (i.e., the item response functions intersect, which means that the order of items in terms of difficulty is not the same for respondents with different levels of the latent trait). The criterion values reported as a result of the check of differential item functioning (DIF) appear to be all lower than 80, which means that no DIF is indicated.

Third, we did not find any violation of the assumption of nonintersecting IRF's for the special needs group. For the control group, a violation of this assumption was found for 12 items in the cognition scale, and one item in the Fine as well as Gross motor scale, see Table 4. The MSP results of the control group data showed that the items were approximately ordered by difficulty, as shown by the item means which were roughly decreasing. We counted the number of Guttman errors, which are the number of item pairs with a 0 on the easier item and a 1 on the more difficult item (Meijer, 1994), to get an impression of the degree to which the items are ordered by difficulty for the control group. The average number of Guttman errors per scale was 15.3 (SD = 14.9) for cognition; 4.6 (SD = 6.7) for receptive communication; 2.9 (SD = 4.4) for expressive communication; 6.2 (SD = 7.4) for fine motor development; and 5.6 (SD = 6.6) for gross motor development.

Fourth, no differential item functioning between the special needs group and the control group was indicated for any of the scales of the Bayley-III. Table 4 shows the criterion values that were the result of the check of equal item step ordering across groups, which are all below 80.

3.1.3. IRT: specific findings

When the assumption of monotonicity appeared to be violated, we checked the corresponding IRF's to locate the cause of the violation.

For cognition, we found a violation of the assumption of monotonicity for four out of the 91 items (numbers 1, 2, 5, and 46) in the control group. The criterion values for these items ranged between 83 and 141. Inspection of the empirical IRFs belonging to these four items revealed that the violations were due to empirical values that were based on a sample of 16 children in the case of items 1 and 2, and 25–27 children in the case of the other two items.

For the receptive communication scale administered to the children in the special needs group, a violation of the assumption of monotonicity was indicated for one item (item 19, "Identifies three objects", criterion value = 109). This item was administered to 22 children. Inspection of the empirical IRF revealed that children with a rest-score (which is the total raw scale score minus the score on this item) between 9 and 17 had a slightly higher average score on this item than children with a rest-score between 18 and 25, see Fig. 2. Inspection of the data matrix revealed that the scores of most children were in line with their rest-scores, and that the violation was caused by deviant scores on this item of approximately four children. For the control group, the assumption of monotonicity appeared to be violated for item 49, the last item in the scale ("Identifies categories", criterion value = 87). Inspection of the IRF and the data matrix revealed that this was caused by the fact that only one child out of the 1063 received a score of 1 on this item, and that this child did not have the highest raw score possible, namely 41.

For the expressive communication scale no violations of the assumption of monotonicity were indicated for any of the two groups.

Fig. 2. IRF of item 19 receptive communication, special needs group. The item response function of item 19 of the receptive communication scale: "Identifying three objects", based on the responses of the children in the special needs group. The violation of the assumption of monotonicity (e.g., the IRF is monotonically non-decreasing) is visible in the figure: between restscore 9–17 and restscore 18–25 the line descends, while an increase would be expected: a higher restscore implies a higher level of receptive communication, hence a higher proportion of positive responses on item 19. The drop is, however, minor.

For fine motor development, the results of the control group showed a violation of the assumption of monotonicity for item 1 (“Hands are fisted”, criterion value = 129). Inspection of the empirical IRF and data matrix revealed that this was caused by the fact that only seven children obtained a score of 0 on this item, and that the raw scores of these children ranged from 5 to 12, while 29 children in the control group obtained a raw score of 4 or less.

For gross motor development, the results of the special needs group on item 66 (“Stops from a full run”, criterion value = 96) appeared to violate the assumption of monotonicity. Inspection of the IRF and the data matrix revealed that this was probably caused by the results of one child, who obtained a relatively low raw score of 55, but did obtain a positive score on item 66. The other children with a positive score on this item all had a raw score of 63 or higher. For the control group, the results for item 1 (“Thrusts legs in play”, criterion value = 195) violated the assumption of monotonicity. This appeared to be caused by the score of one child, who obtained a score of 0 on (the easiest) item 1, but a raw score of 9 on the Gross motor scale (while 60 of the 1066 children in the control group obtained a lower raw score).

3.2. Evaluation form

Table 5 summarizes the results on the main question in the evaluation form. For 40 out of the 69 children in the special needs group, the evaluation form was returned by the test administrator. The test appeared to be suitable in 33 cases, and in 10 of these cases the test administrator reported that this was specifically due to the Low Verbal accommodations. Eight of these 10 respondents reported that the accommodated item instructions were advantageous, and two respondents reported that the accommodated test procedure (e.g., extra emphasis on key words) was advantageous. When the suitability was not caused by the Low Verbal accommodations, the child only had an expressive language problem or scarcely responded to his or her environment, even if gestures were used.

Four respondents reported that the child had not been able to show his/her abilities during the test. Three respondents reported that this was caused by the impairment of the child; two respondents reported that it was caused by behavior, health, or concentration problems of the child. Three of the four children had a severely delayed development. Two of the four children had a pervasive developmental disorder or signs thereof, and one of the other two children had a motor impairment (together with the language impairment).

3.3. Interviews

The interview consisted of six questions. Below, we have summarized the answers to the questions and added, between brackets, the number of times it was reported.

The first question was: “To what extent do you think the Bayley-III-NL is suitable for the target group of your organization?” The respondents evaluated the Bayley-III-NL in general as suitable, especially for children between 2.0 and 2.6 years of age, and children with challenging behavior for whom the SON-R 2.5–7 is not yet suitable (2×). Disadvantages of the Bayley-III-NL are that results are often not accepted for school referrals (2×), and that the instrument does not yield information about scholastic skills (1×).

The second question was: “What is your opinion about the practicality and the added value of the Low Verbal accommodations?” The practicality of the gestures and the extra emphasis on key words is as moderate (1×) to good (3×), and of the shortened sentences good (4×). The added value is moderate (3×) to good (1×) for the gestures; and moderate (2×) to good (2×) for the shorter sentences and extra emphasis on key words. The gestures are clear but give less information to the child than the linguistic instruction does, and hence do not give sufficient support (2×), especially from item 70 onwards (1×). The Bayley-III-NL Low Verbal still contains rather long sentences (1×). The shorter sentences and extra

Table 5
Number of response options chosen to main question in evaluation form: “Has the child been able to show his/her abilities on the test?”.

Response option	N
Yes, caused by Low Verbal accommodations	10
- Adjusted item instructions	8
- Adjusted test procedure	2
- Other, namely...	0
Yes, not caused by Low Verbal accommodations	23
No	4
- Caused by behavior/health/concentration problems child	1
- Caused by impairment child	2
- Caused by behavior/health/concentration problems & impairment child	1
- Caused by characteristics instrument	0
- Caused by environmental factors	0
No answer	3
Total	40

emphasis are not sufficiently helpful in cases of low verbal abilities (1×). The gestures are, however, logical and natural, and provide good support (1×).

The third question was: “Does the Bayley-III-NL Low Verbal have added value, compared to the standard version, for the target group of your organization?” The Low Verbal has “a little” (1×) or “clear” (2×) added value with respect to the test result, and “clear” (3×) added value with respect to the course of the test administration.

The fourth question was: “For which children does the Bayley-III-NL Low Verbal specifically have added value, compared to the standard version?” The Low Verbal is especially suitable for children who do not have too large a delay in their receptive language development, and for whom the SON-R 2.5–7 is not yet suitable. For the other children, the SON-R 2.5–7 is preferred (2×). The Low Verbal is especially suitable in cases of limited concentration abilities (1×), suspected developmental delay (1×), and speech-language problems with little active language (1×).

The fifth question was: “How did the course of the administration of the standard communication and motor scales of the Bayley-III-NL go, and to what extent did the results give a good picture of the development of the children in these areas?” The standard assessment of the communication and motor scales give a fairly reliable picture of the development of the child (3×). The results on these scales were insightful (1×), and the scales are child-friendly (1×).

The sixth question was: “What do you find are the advantages and disadvantages of the Bayley-III-NL Low Verbal when compared to the SON-R 2.5–7?” The advantages are that the test materials are more appealing for the young children, compared to the book-based materials of the SON-R 2.5–7 (5×). The results of the Bayley-III-NL give more information that is relevant in the subsequent support of the child, and in tracking the child’s development (1×). The longer duration causes moments of resistance, which are informative (1×). The Bayley-III-NL has a more suitable abstraction level and is therefore more reliable, especially for the age range of 2.6–3.0 years (1×), and the results are less dependent on earlier experiences with the test materials (1×). Disadvantages mentioned were that the Bayley-III-NL is not a good basis for making a classification of the language impairment for older children in the age range (1×). In addition, it can be difficult to explain that it is cognition that is measured with the Bayley-III-NL, because the items do not have a clear link with scholastic skills (1×). The Low Verbal accommodations do not give sufficient support for children with a severe receptive language impairment (2×).

4. Discussion

The current study focused on the validity and added value of the Bayley-III-NL Low Verbal for the assessment of young children’s development in cases of language impairment. We compared the test results of children with language impairment on the accommodated cognition scale, and on the standard Communication scales and motor scales, to those of children without an impairment on the standard Bayley-III-NL. We asked the test administrators and four developmental psychologists whether the Bayley-III-NL Low Verbal is suitable for children with a language impairment.

The results of the Mokken scale analysis showed that the items meet the monotonicity assumption and that the scales sufficiently distinguish between children with different levels of the latent trait. The check of monotonicity revealed a few items for which the assumption of monotonicity was violated. The violations were minor, however, because they were based on the test results of relatively few children. The fact that no violations were found in the special needs group for the cognition scale could be due to the relatively small sample in the special needs group.

The results of the check of nonintersecting IRF’s in the control group show that the assumption of invariant item ordering does not hold for the cognition scale, even though the US manual indicates that the items were ordered in increasing difficulty (Bayley, 2006a), but it seems to hold for the other scales. The fact that the assumption of increasing difficulty of items might not hold for all scales means that the imputation of scores on the basis of the Bayley-III-NL scoring rules might have biased the data, possibly resulting in an overestimation of the quality of the scales.

No DIF was found between the special needs group and the control group for any of the items in any of the scales. DIF could have been caused by an unwanted influence of the language impairment on the measures of cognition and motor development. Also, DIF could have been caused by the accommodations to the cognition scale influencing the construct that is measured. For example, using short sentences and added emphasis on key words could have altered the memory or attention demands, which are part of cognition. The fact that no DIF was found suggests that the accommodations have not changed the construct that is measured by the cognition scale, supporting the construct validity of the Bayley-III-NL Low Verbal version. Also, it suggests that the Bayley-III-NL Low Verbal contains items that function similarly for children with a language impairment as the corresponding items in the standard Bayley-III-NL for those without a language impairment. If those conclusions are correct, the Bayley-III-NL Low Verbal can sensibly be used for assessing children with a language impairment. Because of the relatively small sample size, those conclusions should be drawn with caution.

The results on the evaluation form reveal that the test administrators found the Bayley-III-NL Low Verbal suitable for a vast majority of the children. For some children, the Low Verbal accommodations played a role in this suitability, but for many others the standard Bayley-III-NL would have been suitable as well. The results of the semi-structured interviews with developmental psychologists confirmed that the Low Verbal accommodations do not have added value for all young children with a language impairment. More specifically, the developmental psychologists reported that the accommodated instrument is especially suitable for children who have a mild or moderate delay in their receptive language development and for whom the SON-R 2.5–7 is not yet suitable. They preferred the Bayley-III-NL Low Verbal for children up until about 3 years of age, and reported that it is also suitable for children with general developmental delay. The respondents reported

that the Bayley-III-NL has clear advantages, when compared to the SON-R 2.5–7, in terms of the suitability of the test materials. In addition, they reported that the unaccommodated communication and motor scales gave a reliable picture of the development of the children.

Taken as a whole, the results mean that the Bayley-III-NL Low Verbal cognition scale, and the standard communication and motor scales, can validly assess the development of young children with a language impairment with a developmental age up to approximately 3.6 years. The Low Verbal accommodations are specifically advantageous for children up to 3 years of age, with or without a more general developmental delay in addition to the language impairment. For older children with a severe receptive language impairment, the SON-R 2.5–7 is more suitable. On a child-by-child basis, it is up to the developmental psychologist to judge which instrument is the most suitable. This recommendation is in accordance with the results of a study into the “Low Motor/Vision version” of the Bayley-III-NL, accommodated for children with a motor and/or visual impairment (Visser, Ruiter, Van der Meulen, Ruijsenaars, & Timmerman, 2013). The Low Motor/Vision version also appeared to be suitable for a subgroup of children in the target group, and the advice with respect to that version was similar: the developmental psychologist must judge which version is the most suitable for a specific child.

A limitation of the current study is that the added value of the accommodations could not be evaluated on the basis of a comparison between test results on the standard and Low Verbal versions of the Bayley-III-NL. This was caused by the fact that it appeared to be impossible to administer the standard Bayley-III-NL to children with a language impairment, because that led to an awkward testing situation as it came across as very unnatural and insensible when the test administrator restrained from using supportive gestures while they were actually needed. As a consequence, we depended on the opinion of the test administrators and developmental psychologists to evaluate the suitability. From a clinical perspective, this opinion is valuable.

A second limitation is the relatively small sample of the special needs group. The effort it took to recruit the children in this sample showed that it would not be feasible to obtain a larger sample. However, a sample of 69 children is rather small as a basis for Item Response analysis. In addition, the scoring rules of the Bayley-III resulted in a situation where some of the items were administered to only a few of the children or even to none of the children at all. The scores were, in these cases, based on the basal and ceiling rules of the Bayley-III. As a consequence, the *H*-values found in this study could, for example, be an overestimation, while the conclusions on the basis of the results of the MSP analyses can only be made for a range of the items per scale. However, it was especially the easier items, those at the low end of the scales, which were not assessed in a sufficient number of children. This was probably caused by the fact that the easier items are meant for very young children, and a language impairment might not yet be identified at a very young age. The items about which we can draw no conclusions thus might not be used very often in clinical practice with children with a language impairment. In addition, the Low Verbal accommodations were relatively little applied to the items in the very low age group (<12 months), because these items did not contain language as a consequence of being suitable for very young children who do not yet have a lot of language.

A third limitation is that the indicated absence of DIF between the special needs and control group could be due to the limited sample size. This might be more likely for minor forms of DIF that would end up only barely affecting the test scores, and thus would not be relevant from a clinical perspective; however, the presence of DIF that remained unidentified in this study cannot be completely ruled out.

A strength of the current study is the fact that the test results of the special needs group could be compared to a very large control group, which was part of the standardization sample in the Netherlands. Another strength was that developmental psychologists who work with young children with a language impairment have been involved in the development as well as the evaluation of the accommodations.

Future research should aim at continuously improving the assessment possibilities for young children with a language impairment. In addition, an appropriate developmental assessment instrument is needed for young children who are deaf. One of the experts interviewed for the pilot study proposed translating the Bayley-III-NL into a sign language version. Another option would be to make an entirely Nonverbal version of the Bayley-III-NL by removing all items that contain language. The target group, however, would be relatively small, because nowadays most children with deafness receive a cochlear implant at a young age. The target group of young children with a language impairment is much larger, and we would like to once again stress the importance of the availability of a valid standardized instrument for developmental assessment of children in this target group. Because early language impairment is associated with problems in other developmental areas in the short as well as the long term (Johnson et al., 2010; McCabe, 2005), it is essential that the development of the children in the target group can be assessed and tracked in different domains. The intention of the current study was to contribute toward reaching this goal.

Acknowledgements

This study is supported by grant 157013002 from The Netherlands Organization for Health Research and Development. The publisher of the Bayley-III-NL, Pearson Assessment and Information BV, supplied the testing materials. We would like to thank the children who participated in this study and their families for their cooperation. We also thank the professionals from the referring organizations for referring the children, for facilitating the test administrations, and for their helpfulness in the interviews. We would also like to thank the test administrators for their efforts in administering the tests for the current study, and in particular Ineke Bezemer, who contributed significantly to the development of the Bayley-III-NL.

Low Verbal and to the pilot research. In addition, we would like to thank the research team at Utrecht University, who perform the standardization research on the Bayley-III-NL, for sharing their research data.

References

- Bayley, N. (1993). *Bayley scales of infant development* (2nd ed.). San Antonio, TX: The Psychological Corporation.
- Bayley, N. (2006a). *Bayley scales of infant and toddler development* (3rd ed.). San Antonio, TX: Harcourt Assessment.
- Bayley, N. (2006b). *Bayley scales of infant and toddler development. Technical manual* (3rd ed.). San Antonio, TX: Harcourt Assessment.
- Botting, N. (2005). Non-verbal cognitive development and language impairment. *Journal of Child Psychology and Psychiatry*, 46, 317–326. <http://dx.doi.org/10.1111/j.1469-7610.2004.00355.x>
- Crais, E. R. (2009). Testing and beyond: Strategies and tools for evaluating and assessing infants and toddlers. *Language, Speech, and Hearing Services in Schools*, 42, 341–364.
- Dodd, B., & Crosbie, S. (2002). Language and cognition: Evidence from disordered language. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 470–490). Malden, MA: Blackwell Publishers Ltd.
- Elbro, C., Dalby, M., & Maarbjer, S. (2011). Language-learning impairments: A 30-year follow-up of language-impaired children with and without psychiatric, neurological and cognitive difficulties. *International Journal of Language & Communication Disorders*, 46, 437–448. <http://dx.doi.org/10.1111/j.1460-6984.2011.00004.x>
- FL State Board of Education. (2010). Exceptional student education eligibility for students with language impairments and qualifications and responsibilities for the speech-language pathologists providing language services, Rule 6A-6.030121..
- Guralnick, M. J., & Conlon, C. J. (2007). Early intervention. In N. J. Roizen (Ed.), *Children with disabilities* (6th ed., pp. 511–521). Baltimore, MD: Paul H. Brookes Publishing.
- Johnson, C. J., Beitchman, J. H., & Brownlie, E. B. (2010). Twenty-year follow-up of children with and without speech-language impairments: Family, educational, occupational, and quality of life outcomes. *American Journal of Speech-Language Pathology*, 19, 51–65.
- Mashburn, A. J., & Myers, S. S. (2010). Advancing research on children with speech-language impairment: An introduction to the early childhood longitudinal study—Kindergarten cohort. *Language, Speech, and Hearing Services in Schools*, 41, 61–69.
- McCabe, P. C. (2005). Social and behavioral correlates of preschoolers with specific language impairment. *Psychology in the Schools*, 42, 373–387.
- Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, 18, 311–314.
- Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, 9, 354–368. <http://dx.doi.org/10.1037/1082-989X.9.3.354>
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, The Netherlands: Mouton.
- Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for windows, a program for Mokken scale analysis for polytomous items*. Groningen, The Netherlands: IEC Progamma.
- Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Newbury Park, CA: Sage.
- Ruiter, S. A. J., Hoekstra, A. T., Van der Meulen, B. F., Lutje Spelberg, H. C., & Nakken, H. (2006). De niet-verbale versie van de BSID-II-NL [The nonverbal version of the BSID-II-NL]. *Tijdschrift Voor Orthopedagogiek*, 45, 406–413.
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Thousand Oaks, CA: Sage Publications.
- Steenis, L. J. P., Verhoeven, J. C. T., & Van Baar, A. L. (2012). The Bayley III: The instrument for early detection of developmental delay. In A. M. Columbus (Ed.), *Advances in psychology research* (pp. 133–141). Hauppauge, NY: Nova Science Publishers.
- Tellegen, P. J., Winkel, M., Wijnberg-Williams, B. J., & Laros, J. A. (1998). *Snijders-Oomen niet-verbale intelligentietest SON-R 2½–7. Handleiding* [Snijders-Oomen nonverbal intelligence test SON-R 2 ½–7. Manual]. (2nd ed.). Göttingen, The Netherlands: Hogrefe Verlag.
- Thal, D. J. (1991). Language and cognition in normal and late-talking toddlers. *Topics in Language Disorders*, 11, 33–42.
- Thurlow, M. L., Elliot, J. L., & Ysseldyke, J. E. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements* (2nd ed.). Thousand Oaks, CA: Corwin Press.
- Van Daal, J., Verhoeven, L., & Van Balkom, H. (2009). Cognitive predictors of language development in children with specific language impairment (SLI). *International Journal of Language & Communication Disorders*, 44, 639–655. <http://dx.doi.org/10.1080/13682820802276930>
- Van der Meulen, B. F., Ruiter, S. A. J., Spelberg, H. C. L., & Smrkovsky, M. (2002). *Bayley scales of infant development – Nederlandse versie [Dutch version]* (2nd ed.). Amsterdam: Harcourt.
- Visser, L., Ruiter, S. A. J., Van der Meulen, B. F., Ruijsenaars, A. J. J. M., & Timmerman, M. E. (2013). Validity and suitability of the Bayley-III low Motor/Vision version: A comparative study among young children with and without motor and/or visual impairments. *Research in Developmental Disabilities*, 34, 3736–3745. <http://dx.doi.org/10.1016/j.ridd.2013.07.027>