The BSID-II-NL: construction, standardisation, and instrumental utility

Selma A.J. Ruiter, Henk C. Lutje Spelberg, Bieuwe F. van der Meulen and Han Nakken

In this article we present normative and psychometric data for the Dutch version of the Bayley Scales of Infant Development – Second Edition. The BSID-II-NL consists of a translation of the original materials and a Dutch standardisation for the mental, motor and behaviour rating scales. The Dutch translation of the BSID-II was made in such a way as to stay as close to the original version as possible. Only two important adaptations have been made: a broadening of the basal and ceiling levels and a different factor structure for the behaviour rating scale. Studies on the psychometric qualities have proved the BSID-II-NL to be reliable and valid. (*Netherlands Journal of Psychology*, 64, 15-40.)

This article introduces the BSID-II-NL and presents the normative and psychometric data for the Dutch version of the Bayley Scales of Infant Development – Second Edition. The addition 'NL' refers to the Dutch translation and adaptation of the test material and the addition of a Dutch standardisation to the original instrument (BSID-II).

The decision to develop a Dutch translation and standardisation of the second edition of the Bayley test, the BSID-II (Bayley, 1993), was made because of the necessity of updating the norms and replacing the outdated materials of the 'old Bayley', the BOS 2-30 (van der Meulen & Smrkovsky, 1983). A second important consideration was that with the release of the BSID-II-NL, the good reputation of the Bayley Scales in the field of early child diagnostics in the Netherlands could be preserved.

In this article the following research question is answered: *Is the Dutch translation and adaptation of the BSID-II a valid and reliable instrument for individual developmental assessment?*

The justification of the standardisation and results from validity and reliability studies with the BSID-II-NL provides us with the answer to this research question.

This article is arranged as follows. First a summarised overview is given of the construction of the BSID-II-NL. The construction phase was concluded by a pilot study that was used to examine the experimental version of the BSID-II-NL. Based on the results of the pilot study and the experiences and observations of the examiners, the final version of the BSID-II-NL was constructed. The most important results and caveats are dealt with. Next we describe the final instrument, a justification of the standardisation study performed on this final instrument and several

University of Groningen
Correspondence to: Selma A.J. Ruiter, Faculty of Behavioural and Social Sciences, Institute for the Study of Education and Human Development (ISED), Grote Rozenstraat 38, NL 9712 TJ Groningen, e-mail: s.a.j.ruiter@rug.nl
Received 7 December 2007; revision accepted 5 December 2007.

studies into its validity and reliability. In the United States extensive studies have been performed on the psychometric qualities of the BSID-II. These studies proved the instrument to be both highly reliable and valid (Matula & Aylward, 1997). The fact that a test is considered reliable and valid in the United States may be a good indication that the same holds for a Dutch translation and adaptation, but this cannot be guaranteed (Drenth, & Sijtsma, 1990). That is why a separate study was conducted in the Netherlands to determine the psychometric qualities of the BSID-II-NL. The article finishes off with a discussion on the use of the instrument in the field of pedagogy and educational sciences.

Construction of the BSID-II-NL

Development of the pilot version

The Dutch translation of the BSID-II (Bayley, 1993) was made in such a way as to preserve the original meaning of the item instruction as closely as possible. Adaptations to the text were only made in two ways: some minimal adaptations were made that were necessary for understanding, for instance translating American measures to the metric system, and some of the possible answers in which American terms like 'sneaker' for shoe or brand names like 'Chevy' for a car were changed to Dutch alternatives. A summarised overview of the information in the original BSID-II manual concerning theoretical background and construction of the new scales was added to the Dutch manual. Finally, based on a literature study (Sarneel, 1999), the decision was made to include broader basal and ceiling rules for the mental and motor scales in the Dutch version.

Publications concerning the BSID-II (e.g. Gauthier, Bauer, Messinger & Closius, 1999; Washington, Scott, Johnson, Wendel & Hay, 1998) show that the division of the mental and motor scales into item sets (belonging to certain age groups) and the accompanying basal and ceiling rules can be the cause of a less accurate estimation of the developmental level of a child. Because of the item sets and the strict basal and ceiling rules, a child is only tested outside the original item set when it performs exceptionally well or exceptionally badly. Therefore, the choice of the initial item set affects the test scores of the child. For this reason the decision was made to require broader basal and ceiling rules for the Dutch version, which provide a broader but also a more accurate picture of the developmental level of a child. A detailed description of the item sets and the basal and ceiling rules of the BSID-II-NL is given further on in this article.

A pilot study was carried out with the experimental version of the BSID-II-NL intended to verify if the translation of the original materials was well phrased and suitable for the target group. Subsequently, it was important to determine if the original arrangement and sequence (based on item difficulty) of the items in the item sets could be maintained.

With the results of the pilot study, the final instrument could be constructed to conduct the Dutch standardisation and the assessment of validity and reliability of the BSID-II-NL.

The pilot study

Sample

The experimental version of the BSID-II-NL was administered to 235 children in the age group of 1 month to 3.5 years. All the children were developing normally, i.e. until the moment of testing the children had not been diagnosed and/or treated for developmental problems. Dutch was the primary language for all the children. Randomised recruitment was conducted via the provincial inoculation administration in Groningen. Table 1 summarises the sample by age group and sex.

Procedure

All tests were administered in the playroom of the Department of Pedagogy and Educational Sciences by either pedagogy graduates or master students. All took part in a brief training in the administration of the BSID-II-NL and were counselled on a weekly basis during the pilot study. In as far as was possible, all children who were registered by their parents were tested if their age corresponded with one of the age groups that had been determined beforehand. For children younger than 24 months a margin of one week younger or older than the exact age was allowed. For children older than 24 months, two weeks difference was allowed. The age groups were selected in a way that ensured that every item would be administered in at least one of the age groups.

Results

The first aspect of this pilot study involved the phrasing and content of the translated test procedure and item instruction. The most important results of the pilot study with regard to test procedure, item instructions and play material of the mental and motor scales and the behaviour rating scale (BRS) are discussed.

Concerning the *test procedure* the examiners found that the order of administration of the items as is noted on the cue sheets (the advised order of the items based on play materials needed and content) is preferable to administering the items in order of difficulty, as on the score form. This cue sheet order should, however, not be compulsory. The directions in the manual with regards to the testing procedure rightly denote that item order and the order in which mental and motor scales are administered depend on the child and on the testing situation.

Table 1 **Number of children per age group and sex in the sample.**

Age group	*Girls*	*Boys*	*Total*
1	12	9	21
3	16	15	31
4	10	9	19
7	9	22	31
11	15	13	28
15	15	16	31
18	10	12	22
24	16	15	31
40	10	11	21
Total	113	122	235

Examiners' experiences with the BSID-II-NL and a further examination of the pilot version have led to improvements in the *item content* (correction of errors, addition of missing information) and phrasing (correcting grammatical errors and ambiguities).

From experience with the *play material* it became clear that, among other things, the stimuli booklet and the doll broke easily. It was advised to replace these materials with more solid specimens. Also, it was found that the play materials in different testing kits were not the same in all instances. This was true for instance for the bunny, the car and the booklet for reading to the children. The BSID-II-NL is a standardised test and therefore care should be taken to create conditions as similar as possible for each test administration, which also means providing the same play materials. Apart from the remarks mentioned above, the play materials were generally judged to be attractive and suitable for the target group and skill to be measured.

The second aspect of the pilot study concerned the division of the BSID-II-NL into age groups based on the original (US) item difficulty. The most easy (first) item in the age group is performed correctly by 95% of the children and 10% of the children receive a positive score on the most difficult (last) item. The *p* values (item difficulty) of Dutch children in the pilot study deviate from these values. In all age groups we found lightly to strongly fluctuating *p* values. While it is true that the last items administered in an age group have lower *p* values, a 95 to 10% difference could not be found. Further analysis of these results gave no cause for changing (the item compilation of) the age groups or the order of items in these age groups. A calculation of correlation between the item numbers and *p* values per item set gives a sufficiently negative correlation (table 2). This table shows that the items are progressively more difficult; the more negative, the better the correlation is between item number and difficulty level. The small fluctuations in *p* values are, in light of the limited sample, not enough reason to change the item order and by doing so the content of the item sets.

No special attention was paid to the behaviour rating scale (BRS). The reasons for not doing this were that the clear phrasing of the questions and possible answers gave no cause to expect a need for modifications in the BRS based on the results of the pilot study and that the limited size of the pilot study did not make it possible to perform a factor analysis to investigate the original (American) factor structure of the items in separate age groups. Only a few corrections to the text were made based on experiences of the examiners in the pilot study.

Conclusion

Except for the broadening of the basal and ceiling rules and some changes in the wording, no significant changes were made to the testing procedure or item instructions of the three scales. No cause was found to adapt the original age group division or item order in the final version of the Dutch version of the BSID-II. In the standardisation study, after each test all questions on the BRS are to be answered by the examiner. The results collected in this way will be used as the basis for a factor analysis to investigate the original factor structure and adapt this to the Dutch situation if needed.

Table 2	Correlation between item numbers and p values per item set on the mental and motor scales.	
Item set	*Mental scale*	*Motor scale*
1	-0.76	-0.60
3	-0.52	-0.76
4	-0.65	-0.79
7	-0.72	-0.72
11	-0.82	-0.83
15	-0.74	-0.51
18	-0.83	-0.73
24	-0.80	-0.55
40	-0.62	-0.71

After the results of the pilot study were processed, the final instrument was put together. The content and characteristics of the BSID-II-NL are summarised below.

Final version of the BSID-II-NL

A summary is given of the content and characteristics of the final instrument (a comprehensive description is provided in Van der Meulen, Ruiter, Lutje Spelberg & Smrkovsky, 2002). For reasons of legibility only BSID-II-NL is mentioned. Apart from some (aforementioned) differences, the text also applies to the original (US) BSID-II.

Three scales

The Bayley Scales of Infant Development have been constructed in such a way as to present the child with tasks and situations that are expected to be of interest to the child. The instrument comprises three scales: the mental scale (178 items), the motor scale (111 items) and the behaviour rating scale (30 items). The items on the mental and motor scales are scored dichotomously (positive/negative), answers to the questions in the BRS are awarded a score from 1 to 5.

Most of the items in the mental and motor scales are task items. In these items, the examiner must try to produce the desired behaviour in the child by presenting stimuli. Some of the items do not use stimuli, they have to be observed spontaneously. The BRS contains two questions for the parent/caregiver in which the parent/caregiver can indicate if the test accurately reflects the child's skills. Next to these questions, the scale comprises a series of items that allow the examiner to judge the behaviour of the child during the test.

The *Mental Scale* consists of 178 items that measure the child's cognitive skills. The mental scale consists of items related to the processing of visual and auditive information, eye-hand coordination, imitation, language development, memory and problem solving. For an example, see appendix 1.

After administrating the mental scale a raw score is calculated (RS_{MS}) and a mental development index (DI_{MS}) is determined. A 90% reliability interval is also provided, and it is possible to determine how the score should be classified: seriously delayed, delayed, normal or accelerated development.

The *Motor Scale* consists of 111 items that measure skills related to gross and fine motor control, including movements like rolling, crawling, standing, walking, running and jumping. This scale also tests fine motor manipulations, such as eye-hand coordination, adjusted use of writing materials and imitation of hand gestures. For an example, see appendix 1. For the motor scale a raw score (RS_{MR}), a motor development index (DI_{MR}), a reliability interval and a classification are also determined (as in the mental scale).

The *Behaviour Rating Scale* consists of 30 items, the first two of which are questions for the child's parent. This scale is administered after the mental and motor scales and is intended to measure the behaviour of the child during the test. Behaviour and disposition of the child are judged, based on level of alertness, adaptation to surroundings and quality of motor skills. For an example, see appendix 1. Scoring all items on a five-point scale and describing the score on the scale improves the reliability of the scale and facilitates scoring and interpretation. Items are grouped in three age groups: 1 to 5 months, 6 to 12 months and 13 to 42 months. Scores on the items of the behaviour rating scale are differentiated by age groups into factors and total score. For each age group three behavioural factors can be determined and a total score. These factor and total scores are expressed as percentages. A description of the original factors is of little use here due to the still unknown results of the factor analysis of the results of the data from the standardisation study. To determine whether the American factor structure also applies to the Dutch situation, a large amount of data is needed. The final structure of the behaviour rating scale is described later in this article.

Test procedure

Bayley's opinion on the structure and the process of child development has allowed the test to be administered flexibly within a standardised procedure. In the manual, specific directions are given for presenting items to the child. The order and the speed with which items are administered depend on a combination of factors, including the age of the child, its disposition and the success rate on related items. Despite the flexible administration procedure, the BSID-II-NL can be characterised as a 'power test'. A power test is a test in which the items are presented in order of increasing difficulty. This means that the difficulty of items that are administered in a certain age group ranges from very easy to very difficult.

For reasons of time limitations (and frustration limitations for the child) and uniformity, item sets were classified according to a broad range of items correlated to the child's age. The item sets show some overlap; the most difficult items for 9-month-old children are for instance the easiest ones for 13-month-old children. The item sets are arranged in such a way as to make the content range of the item set large enough to fix the basal and ceiling values within a single item set, and also to establish sufficiently differentiated scores. Because of the classification of items into age groups, ceiling and basal rules are established instead of cut-off rules. These rules indicate when the lower and upper limits have been reached. If a limit is not reached, the extra items of related item sets will be administered, which are not included in the original item set. The Dutch version of BSID-II has adapted basal and ceiling rules. Instead of at least five positive items within the item set on the mental scale, a child performing the BSID-II NL should score at least eight positive items to reach the basal level in the item set. In the Dutch version the child has to perform better to stay within the item set, about a third of all the items have to be scored positively. If a child happens to perform well in one area, say building blocks, this one skill is not enough to reach the basal level. The ceiling level has also been adjusted so that it is easier to administer items outside the item set. In the original version, the mental scale has a ceiling level of three negative items, in the Dutch version a maximum score of five negative items is a reason for administering items from a higher item set. For the motor scale the basal and ceiling levels have been adjusted to five positive items and three negative items, respectively. By extending the limits, an item set will become less rigid; the child is less likely to get 'stuck' in the initial item set. The child has to achieve more positive scores to reach the basal level, thus having to return more often to a lower item set. In order to reach the ceiling level the child must have more negative scores. By doing this, the item set at the start will exert less influence on the final test result. By going back or moving up from an item set the examiner is given a broader picture of the child's development. Table 3 provides an overview of the original (US) and adjusted (NL) basal and ceiling rules.

Conclusion

The BSID-II-NL is a translation of the original BSID-II and only one significant change is made prior to the standardisation study: the use of adapted basal and ceiling rules for the mental and motor scales. Other than this, the original instrument has remained unchanged both in structure and in content. Based on the results from the standardisation study, additional changes will be made. Naturally, this entails the development of Dutch norms, but also determination of the item difficulty and final factor structure of the behaviour rating scale.

Table 3 **Comparison of the original (US) and adjusted (NL) basal and ceiling rules.**

	Original (US)		*Adjusted (NL)*	
	Mental scale	*Motor scale*	*Mental scale*	*Motor scale*
Basal is reached when:	≥5 items positive	≥4 items positive	≥8 items positive	≥5 items positive
Ceiling is reached when:	≥3 items negative	≥2 items negative	≥5 items negative	≥3 items negative

Dutch standardisation

To provide valuable information regarding a child's developmental level, it is necessary to compare test results of a child with results of children in its age group, but also to compare test results between different ages and to compare with results achieved on comparable tests. By converting raw scores into standardised norm scores, age is taken into account. Raw scores were converted to normalised standard scores with a mean of 100 and a standard deviation of 15. Such scales are used for most development and intelligence tests, enabling us to compare the scores.

Sample

The standardisation study was based on the test results of 1909 Dutch children. The study was conducted at four centres: the Department of Pedagogy and Educational Sciences of the University of Groningen (UoG), the Department of Medical Psychology of the University Medical Centre Groningen (UMCG), the Department of Paediatrics – Neonatology of the Sophia Children's Hospital, Erasmus Medical Centre Rotterdam (Sophia) and the Department of Child Psychiatry of the University Medical Centre Utrecht (UMC Utrecht). It involved children experiencing normal development, i.e. until the moment of testing the children had not been diagnosed and/or treated for developmental problems. Dutch was the primary language for all children.

Initially, in the set-up of the standardisation study, a division was made into ten age groups, with the groups selected in a way as to assure that all age groups would have sufficiently differentiated scores. Later, when data from other centres were made available, the number of age groups was expanded. In each centre, recruitment was conducted in a different way. The sample at the UoG was constructed with the help of the Provincial Inoculation Administration. The main reason for contacting parents through this organisation was that all children are registered here in connection with the national inoculation programme. The group of 18-month-old children was selected differently: this group participated in a study at the University Medical Centre Groningen into the effects of polychlorinated biphenyl (PCB) levels in mothers on the mental and motor development levels of the 18-month-old child. The UMCG group also included Dutch speaking children participating in a study into the effects of breastfeeding or bottle feeding on the developmental level at approximately 18 months. Ultimately, children who were born after a pregnancy without complications, and without having been diagnosed or treated for developmental problems until the moment of testing, were tested between the ages of 17 to 24 months. All children in de Groningen sample lived in the city or province of Groningen. The children in the Sophia group from Rotterdam also participated in a national study into the effects of PCB levels in mothers on the psychomotor and mental development levels of the 18-month-old child. For this group of children the characteristics mentioned above also apply. The test subjects presented by the UMC Utrecht participated in a study into the effects of hormones in the prenatal and neonatal period on the development of the brain and development of behaviour, cognition and physiology. These participants had been born in normal single child pregnancies, where an amniocentesis had been performed because of age (96%), medical history (3.3%) or other medical reasons (0.7%). This produced a group of children aged between 13 and 14 months. It should be noted that a small number of children in the UoG sample are present in the sample more than once (in different age groups). The children in the Sophia group sample were all tested when they were 3 months, 7 months and 18 months old. Because the children all met the inclusion criteria and because of the large numbers of children in these age groups, we expect them to reflect the population in the Netherlands in these age groups. Besides this, we do not expect any measurement effect as a result of the young age of the children, large intervals between two test administrations and because of the structure of the BSID-II-NL itself. The division of the items into separate age groups (or intervals) enables the clinician to administer a completely new set of items when retesting the child after 3 or 4 months. Table 4 gives an overview of the number of children per age group and sex in the standardisation study. From the table it is apparent that the number of subjects per age group differs. These sometimes large differences in numbers of children per age group and the unequal intervals between the age groups are largely compensated for by the standardisation method. This is because the norm tables have been constructed through a procedure that is based on the scores of all age groups together. For an explanation of the normalisation method, see later in this article.

Table 4	Number of children per age group (in months) and sex.																	
Age	*1*	*3*	*4*	*7*	*11*	*13*	*14*	*15*	*17*	*18*	*19*	*20*	*21*	*22*	*24*	*36*	*40*	*Total*	*%*
Girls	17	195	16	154	23	21	9	21	11	189	167	22	20	21	29	11	19	945	49.5
Boys	13	191	12	163	16	23	13	24	17	193	164	37	26	19	24	13	16	964	50.5
Total	30	386	28	317	39	44	22	45	28	382	331	59	46	40	53	24	35	1909	100

Because the majority of children from the city and province of Groningen live in 'white' areas, the number of children with an ethnic background is expected to be low. Because of the more culturally mixed living areas in Utrecht and Rotterdam, this may be different for the children in the UMC Utrecht and Sophia Rotterdam groups. Table 5 gives the education level of the parents of the children participating in the standardisation study. These data are compared with data from the Dutch Census Bureau (Statistics Netherlands, the CBS) for the year 1999. This is the year in which the gathering of standardisation data was started. The age group of 30 to 34 years was chosen since it was assumed that this interval closely approximates the average ages of parents of children in the sample.

Because of differences in coding between the different study centres, a further differentiation in education levels is not possible. The CBS data from 1999 have therefore been divided into the same three categories. It is clear from table 5 that the percentages of parents participating in the standardisation study in different education levels closely match the levels of the total population of the Netherlands. A goodness or fit test (χ^2 test) confirms this. There seems to be a dependency between education level and the type of sample (BSID-II-NL versus CBS) both for men, χ^2 $(2, n = 1909) = 13.20$, $p < 0.01$ and women, χ^2 $(2, n = 1909) = 7.29$, $p < 0.05$. However, effect size indices for χ^2 tests for contingency tables were small (see Cohen, 1988, p. 224-225) both for men ($w = 0.08$) and women ($w = 0.06$).

Procedure

During testing, conditions were made as similar as possible for each child. All tests were administered in specially equipped rooms in the study centres. Only a few of the tests were administered at home, under conditions closely approximating those at the centres.

The tests were administered by well-trained and experienced examiners, usually in the presence of a parent/caregiver. In total, 37 examiners participated in the project, divided over the four study centres. By extended training and monitoring by the project leader during the study we tried to minimise the effect of variations in the behaviour of the examiners. The tests of children in Rotterdam and Utrecht were administered by experienced remedial educationalists and psychologists with extensive experience in testing young children. In the case of the children in the Groningen group, the tests were administered

Table 5	Education level of parents in the standardisation sample vs. population averages (provided by the CBS) in percentages.			
*Education level**	*Standardisation sample BSID-II-NL*		*CBS data (1999; 30-34 years)*	
	Male	*Female*	*Male*	*Female*
Low	26.1	22.5	27.9	25.3
Middle	40.8	49.3	43.3	47.5
High	32.3	27.7	28.7	27.2

*The education level of parents has been subdivided into three groups: Low = primary school, technical school, etc. (in Dutch: basisonderwijs, LBO, VBO, MAVO), Middle = O levels, A levels, etc. (in Dutch: HAVO, VWO, MBO), High = higher education (in Dutch: HBO, WO).

by third-year pedagogy students who had received an intensive, two-day training in using the test instrument. After this training they first assisted more experienced examiners and/or video tapes of the first test administrations were made. Those tapes were discussed extensively. During the test period, weekly evaluation sessions were held to compare notes and experiences. These regular evaluation sessions made sure that examiners were constantly aware of their behaviour during the test administrations and of the importance of strictly maintaining the test procedures and item instructions.

Results

After administering the BSID-II-NL, raw scores for the mental and motor scales are calculated and converted to normalised standard scores by means of norm tables, called Development Indices (DI). The behaviour rating scale produces a number of factor scores and a total score. The raw scores of the BRS are converted as well by means of norm tables, only this time in terms of percentiles (as for the original BSID-II).

Mental and motor scales

Table 6 lists the raw scores on the mental and motor scales for the age groups included in the standardisation study.

The norm tables for the mental and motor scales are constructed by a method that uses a fit procedure based on the score differentiation of all age groups together, a method described by, among others, Snijders, Tellegen and Laros (1988) and Van Eldik (1998). The basic assump-

Table 6 **Overview of mean raw scores (mean RS) and standard deviation (SD) per age group on the mental and motor scales.**

	Mental scale			*Motor scale*		
Age (in months)	*Mean RS*	*SD*	*N*	*Mean RS*	*SD*	*N*
1	11.47	3.32	30	11.67	2.17	30
3	31.42	3.20	386	22.23	2.31	386
4	42.64	2.79	28	25.54	1.82	28
7	63.44	3.11	317	39.16	4.00	317
11	78.62	3.01	39	58.31	2.75	39
13	88.79	4.21	43	60.42	3.39	43
14	92.59	5.75	22	63.55	3.50	22
15	97.49	4.87	45	67.31	7.83	45
17	109.21	6.69	28	75.68	2.55	28
18	109.87	6.05	382	75.20	3.17	381
19	113.02	6.49	331	76.38	2.72	331
20	117.36	6.33	59	78.98	3.58	55
21	119.82	8.70	45	80.04	3.69	45
22	124.55	6.97	40	82.15	3.22	40
24	129.94	6.98	52	83.23	3.72	53
36	156.54	4.21	24	100.13	4.48	24
40	163.06	5.76	35	104.11	2.63	35
Total			1906*			1902*

Mean RS = mean raw score, SD = standard deviation, n = number of children. * The totals do not add up to 1909 because both scales were not administered to all children.

tion is that in the test population changes in developmental characteristics change systematically with age. This method was selected for two reasons. First of all, the number of test subjects per age group was too small to determine the norms in the classical way, i.e. separately per age group with sufficient reliability. The limited sample size for the norm groups increases the chance of random fluctuations, especially at the ends of the score range. The second reason was that the fit procedure produces a regression equation that enables the construction of norms for the empirical age groups, but also for the (intermediate) age groups that were not studied.

First, temporary normalised standard scores were determined per age group from the raw scores using a non-linear transformation as described by Lienert (1961, p. 336-344). First the cumulative frequency distribution of raw scores for each age group in the standardisation sample was determined, after which the normalised *z* scores for these cumulative proportions are determined (with a continuity correction). These *z* values are then transformed to a distribution with mean = 100 and standard deviation = 15. As stated, these temporarily normalised standard scores are not very reliable because of the sometimes small number of test subjects per age group. For this reason we performed a gradual non-linear regression analysis for a combination of all age groups, using the temporarily normalised standard score as dependent variable and age and raw score as independent variables. This procedure produced a non-linear regression equation that could then be used to calculate final standardised scores based on raw score and age.

For the motor scale it was not possible to get a good fit for all ages combined. After inspecting the development of the raw scores with age, it was decided to perform two separate analyses: one for the age groups of 1 to 13 months, and one for the age groups 7 to 42 months. In doing so it was possible to get a good fit; this means that the multiple correlates between the final standard scores and the preliminary standard scores are sufficiently high and the number of predicting variables is as low as possible. If needed, corrections were made for not complying with the requirements set by Laros and Tellegen (1991, p. 43): for each specific raw score the final standardised score must decrease as the age increases and per age group final standardised scores have to increase as the raw score increases. Using the final norm scores, norm tables were constructed for the entire age range. Because of the rapid development of children younger than 12 months, the norm tables list norms per six months for this group and per full month for children aged 12 to 42 months. For the regression equations of the mental and motor scales, see appendix 2.

Behaviour rating scale

Using the Dutch test data a principal component analysis was used to determine if the original (American) factor structure and division into age groups could be maintained in the BSID-II-NL. It was found that this structure and division could not be maintained. Only the correlation between the scores on a group of (motor) items of the behaviour rating scale and the factor 'motor quality' corresponded to the American results. The results of this component analysis then led us to perform a second component analysis, this time using two factors. The structure of the 'factor loading' found in this way confirmed that based on the Dutch data, the behaviour factor structure is less differentiated than the American structure. The number of age groups has been reduced too, from three to two and the number of specific behaviour factors for each age group is two, one less than in the US version. For the BRS, percentiles were determined for the raw scores. Because we standardised for combined age groups (with sufficient numbers of children) a fit procedure was not necessary. The percentages were calculated directly from the cumulative proportions, corrected for continuity.

Developmental indices

Based on the standardisation method described above, norm tables were constructed. Using the norm tables, raw scores in the mental and motor scales can be converted to developmental indices (standard scores). These show how a child is developing compared with contemporaries, but also compared with children of other ages. The developmental indices for both scales have, just as for the IQ, a mean value of 100 and a standard deviation of 15, with a lower limit of 55 and an upper limit of 145. Using internal consistency, 90% reliability intervals for the developmental indices were determined for each age group (in this case 'probability intervals' as defined by Evers, van Vliet-Mulder & Groot, 2000, p. 1407). For the age groups in the study these were based on the reliability coefficients that were found; for other age groups they were based on interpolated reliability coefficients.

Developmental age equivalents

The mental and motor capabilities of a child can also be expressed in terms of 'developmental age equivalents'. This expresses the score in terms of the age at which a child with a developmental index of 100 (being the mean for that age) would (on average) achieve the measured raw score. With some care these developmental age equivalents can be calculated for children who are too old to apply the norm tables, but achieve a score that corresponds to the mean score of younger children, or to children who are younger than 42 months, but achieve a score that is outside the norm tables. For this last group of children it is possible to add a (in this case more significant) developmental age equivalent determination to the developmental index < 55.

Conclusion

Because of the standardisation method that was used, full use was made of the large number of children in the standardisation sample. By incorporating the scores of all age groups into a continuous function of age, it becomes possible to determine norms based on the test data of all children combined. This method has many advantages, but in this case it is a necessity, because of the influence of sample variations in the age groups with small numbers of children and because of the uneven distribution of children over age groups. The smaller the number of children, the larger the influence that random and/or extreme scores exert on the mean score of the group. By using the results for all age groups in the standardisation of the BSID-II-NL, it is possible to compensate for the influence of random fluctuations by using scores of children in adjoining age groups. This eliminates the need for large numbers (at least 100) of children in evenly distributed age groups.

In the case of the BSID-II-NL, all standardisation data had to be gathered in a relatively short period and with very limited means. The size of the study in combination with time limitations and the limited means available were not conducive to selecting the sample. The time limitations were caused by the desire to replace the BOS 2-30 as soon as possible because of its outdated norms and the unavailability of the test material. Because of this it was decided to release the test in parts. This led to a first publication of the manual and standardisation, with a second publication containing the technical foundations following later.

With regard to the behaviour rating scale, the BSID-II-NL has a different factor structure. It was not possible to sufficiently substantiate the original factor structure using the data from the Dutch standardisation study. The factor structure in the Dutch version is less differentiated: the number of age groups was reduced from three to two and the number of specific behaviour factors is two for each age group, one less than in the US version.

Instrumental utility

In the United States extensive studies have been performed on the psychometric qualities of the BSID-II. These studies proved the instrument to be both highly reliable and valid (Black & Matula, 2000). The fact that a test is considered reliable and valid in the United States may be a good indication that the same holds for a Dutch translation and adaptation, but this cannot be guaranteed (Drenth & Sijtsma, 1990). That is why a separate study was carried out in the Netherlands on the psychometric qualities of the BSID-II-NL. To examine the instrumental utility, we conducted several reliability and validity studies. We assessed internal consistency, test-retest reliability, stability and inter-rater reliability. To assess construct validity, the relationship was examined between the BSID-II-NL and comparable widely used Dutch instruments. Finally, we discuss the relation with background variables.

Sample

To assess the instrumental utility of the BSID-II-NL, three different samples were analysed. To determine internal consistency, the data from the standardisation sample were used: sample 1 ($n = 1909$). For assessing test-retest reliability, inter-rater reliability and construct validity, sample 2 was constructed by sending a letter to all the parents of children participating in the UoG standardisation sample, asking them to contact us if they were willing to participate in a second test with their child. This provided us with a group of children, some of whom had also participated in the standardisation sample. We added subjects to this sample by inviting (mostly younger) siblings of these children to participate. This sample 2 consists of 168 tested children between the ages of 3 and 42 months (Zwart, 2004). The only criteria for inclusion were that each child should be healthy and between 0 and 42 months of age. On the basis of the age level of the children different subsamples were constructed to be used in separate validity and reliability studies. Some of the children in sample 2 took part in more than one study. For example, some children were examined twice with the BSID-II-NL (test-retest reliability) and one of the administrations was observed by an inter-rater (inter-rater reliability). The subsamples described for each study belong to sample 2. To assess stability (long-term reliability), sample 3 was used. The third sample consists of children ($n = 62$) who were taking part in the UMCG study into the effects of PCB levels. These children were tested at 18 months of age (and included in the standardisation sample) and again at 30 months of age. Table 7 shows the division of the children in the different sample studies.

Procedure

All children in sample 2 and 3 were tested twice. Some were tested twice with the BSID-II-NL as part of the reliability studies, others were tested with the BSID-II-NL and either with the BOS 2-30, the SON 2½-7, or the GOS 2½-4½ to assess construct validity. From here on, we will refer to such a combination as a test combination. All the tests were administered by well-trained and experienced examiners, usually in the presence of a parent/caregiver. Care was taken to ensure that both tests were administered under circumstances that were made as similar as possible; if

Table 7 Characteristics of the 'instrumental utility' sample for the different test combinations.

			Age in months		*Sex*	
Test combination	*Sample*	*N*	*Mean*	*Range*	*Boy*	*Girl*
Reliability						
Internal consistency	1	1909	See table 8			
BSID-II-NL test-retest	2	34	20.8	4.0-39.2	13	21
BSID-II-NL inter-rater	2	35	20.7	3.0-42.0	20	15
BSID-II-NL stability	3	62	18.2*	18-19*	36	26
Construct validity						
BSID-II-NL vs. BOS 2-30[1]	2	43	17.4	3.0-30.0	19	24
BSID-II-NL vs. SON 2½-7[2]	2	28	37.2	30.0-42.0	11	17
BSID-II-NL vs. GOS[3]	2	28	33.8	30.0-41.0	17	11

At second measurement (30 months): 30.44 (30-31). [1]Bayley Development Scales 2-30 months (Van der Meulen & Smrkovsky, 1983). [2]Snijders-Oomen Non Verbal Intelligence Test 2½-7 years (Tellegen, Winkel, Wijnberg-Williams & Laros, 1998). [3]Groningen Development Scales 2½-4½ years (Neutel, Van der Meulen & Lutje Spelberg, 1996).

possible, children were tested at the same time of day and in the same room. The order in which the tests were administered was varied as much as possible. Some children were tested with the BSID-II-NL first and then with one of the others, other children were tested with the other test first. This was systematically varied.

Results

Reliability
There is no single measure for the reliability of a test. By assessing different forms of reliability, an indication of general reliability is given. For the BSID-II-NL, internal consistency, test-retest reliability, long-term stability and inter-rater reliability were determined.

For the BSID-II-NL, *internal consistency* was determined based on the data from the standardisation study (sample 1). To determine internal consistency, the lambda-2 coefficient (Guttman, 1945) was used. According to Ten Berge and Zegers (1978), the lambda-2 coefficient is a less-known but better estimator of the lower limit of reliability than the more often used alpha coefficient of Cronbach (1951). Ten Berge, and Zegers (1978) state that among a number of ways to calculate the lower limit of reliability, lambda-2 is always greater or equal to the alpha coefficient and therefore is at least as close to reliability in the population as the alpha. For this reason and because of the simplicity of the equation and its long existence, Ten Berge, and Zegers (ibidem) advise using the lambda-2 coefficient to determine internal consistency.

Evers, Van Vliet-Mulder & Groot, (2000) indicate in their 'Documentation of Tests and Test Research in the Netherlands' that when a test is used for selection purposes, a reliability of higher than 0.90 is characterised as good, between 0.70 and 0.80 as medium and lower than 0.70 as insufficient. Van Eldik (1998), however, remarks that in the practice of constructing tests, reliability coefficients higher than 0.80 are often deemed satisfactory. Although the mental and motor scales show low coefficients in a number of age groups (mental: 1, 3, 4, and 40; motor: 1, 3, 4, 7 and 11) the mean internal consistency coefficients are acceptable. No age groups were excluded in these scales, especially to guarantee the international equivalence of the BSID-II editions and because it is not advisable to exclude a specific age group (e.g. the motor scale for 4-month-old children) that is not at the ends of the age range. The low internal consistency of some age groups is expressed in broad reliability intervals. This means that a score on a (fictitious) immediate re-test can vary substantially.

To examine *test-retest reliability,* the BSID-II-NL was administered twice within two weeks to 34 children for the mental scale and 31 children for the motor scale. The age of the children ranged from 4.0 to 39.2 months, with an average age of 20.8 months. The mean developmental indices, standard deviation of the results and the correlations between developmental indices are listed in table 9. The developmental indices were used for calculations so that age is not a factor in the analysis.

Table 8	Internal consistency calculated with lambda 2 per age group over administered items.	
	Lambda 2	
Age group	*Mental scale*	*Motor scale*
1	0.65	0.54
3	0.67	0.66
4	0.51	0.26
7	0.58	0.80
11	0.52	0.80
13	0.74	0.88
14	0.85	0.87
15	0.91	0.75
17	0.84	0.76
18	0.87	0.72
19	0.87	0.87
20	0.94	0.80
21	0.91	0.80
22	0.96	0.84
24	0.88	0.77
36	0.81	0.81
40	0.86	0.65
Mean	0.79	0.74

Table 9	Test-retest reliability: correlations and characteristics of the standard scores.					
		Standard scores				
		First testing		*Second testing*		
	Correlation	*Mean*	*SD*	*Mean*	*SD*	*n*
Mental scale	0.75*	108.18	11.90	113.09	14.56	34
Motor scale	0.80*	99.13	18.04	101.23	18.40	31

* Correlations significant with an alpha of 0.01 (two-tailed).

The correlation between first and second testing is acceptable for both the mental scale and the motor scale. Table 9 shows that the mean developmental indices increase between the first and second testing, especially on the mental scale. Discrepancies in scores over this short period of time may be explained by the combination of maturation and, particularly for children 12 months and older, learning effects. The results on the motor scale closely match the population mean. The results on the mental scale, however, clearly differ. Children in the sample score about 8 to 13 points higher in the first and second test administration, respectively. The standard devi-

Table 10 Test-retest reliability: correlations of raw scores for BRS and characteristics of the standard scores,

		Standard scores				
		First testing		*Second testing*		
	Correlation	*Mean*	*SD*	*Mean*	*SD*	*n*
BRS adaptation	0.49*	74.26	6.67	74.37	7.21	19
BRS motor	0.83**	41.55	5.29	41.50	5.82	20
BRS total	0.67**	113.70	11.54	115.85	17.76	20

* Correlations significant with an alpha of 0.05 (one-tailed). ** Correlations significant with an alpha of 0.01 (one-tailed). BRS adaptation = adaptation to environment.

ation of the first administration of the mental scale also differs. This can be explained because of the small sample that allows extreme values to exert more influence on the mean test score.

Test-retest reliability was also calculated for the behaviour rating scale. The BRS was filled in twice for 20 children. The scale 'adaptation to the environment' was filled in twice for 19 children. The item 'alertness' was not included in the calculations of the correlations. 'Alertness' only applies to children in the 1 to 5 month age group. There were too few children participating in the study in this age group to make calculations of correlations of this item relevant. The correlations (table 10) of the BRS were calculated from the raw scores.

The reliability (stability) of the BRS is lower than that of the mental and motor scales. This is not unexpected. Behaviour is, when compared with mental and motor abilities, more sensitive to environmental influences (Bayley, 1993). Besides, the completion of the BRS allows a certain measure of subjectivity. This subjectivity is caused by the content of the items and the way they are scored. The items are scored on a scale of 1 to 5 instead of dichotomously, as in the mental and motor scales. Also, items are not as extensively described as in the mental and motor scales, leaving more room for personal interpretation, necessitating more care in interpreting the data.

To examine *inter-rater reliability,* for 35 children, an independent second examiner was present during testing and also (independently) scored the test. The age of the children in question varied from 3.0 to 42.0 months, with an average of 20.7 months. To expand the available data for inter-rater reliability, a number of test administrations were videotaped. These tapes were later scored by examiners. The average developmental indices and the correlations between the indices are listed in table 11.

The inter-rater reliability of the mental scale is sufficient, and of the motor scale close to sufficient. The difference in reliability on the mental and motor scales can be attributed to the fact that for a correct scoring of some motor skills, one must physically manipulate the child. The inter-rater does not have this possibility. In individual cases this can cause differences. When the quality of the videotaping of the test administration was inadequate, the test was excluded. This explains why not 35 but only 27 children are listed for the motor scale.

Table 11 Inter-rater reliability: correlations and characteristics of the standard scores.

		Standard scores				
		Examiner		*Inter-rater*		
	Correlations	*Mean*	*SD*	*Mean*	*SD*	*n*
Mental scale	0.81*	102.40	16.89	100.80	14.16	35
Motor scale	0.77*	92.96	15.59	89.74	14.63	27

* Correlations significant with an alpha of 0.01 (one-tailed).

Table 12	Stability: correlations and characteristics of the standard scores.					
		Standard scores				
		Aged 18 months		*Aged 30 months*		
	Correlation	*Mean*	*SD*	*Mean*	*SD*	*n*
Mental scale	0.58*	97.92	13.36	98.94	11.04	62
Motor scale	0.33*	89.13	10.36	95.18	16.66	61

* Correlations significant with an alpha of 0.01 (one-tailed).

In determining the *stability* (or long-term reliability) of the BSID-II-NL, data were used from children who took part in the standardisation study and were later tested again with the BSID-II-NL as part of a follow-up study into the effects of PCB and dioxin exposure on the mental and motor development of young children. The first test was administered when the children were 18 months old. The second was administered at the age of 30 months (table 12).

Since very young children were involved and the interval between tests was relatively large, the expectation was (based on earlier studies, Bayley 1949; Siegel, 1981; Molfese & Acheson, 1997) that correlations would be low on both scales. For the motor scale, this expectation was found to be true. The correlation between results on both tests, although significant, is low (0.33). The prediction value on the mental scale at 18 months for the results at 30 months, however, is substantial (0.58), when age and interval are taken into account. On the motor scale, scores for the second test administration are significantly higher.

Construct validity

For determining the construct validity, the BSID-II-NL was compared with several other Dutch development tests. The choice of instruments to use in the comparison was based on the goal of the test (measuring the mental development and/or motor development), the age range of the test (which had to have at least some overlap with the BSID-II-NL), the structure of the test (individual and standardised) and standardisation and psychometric qualities (based on data from Dutch children). The tests that qualified, based on these criteria, were the BOS 2-30, the SON 2½-7 and the GOS 2½-4½. The validity study was based on test results for 99 children with an average age of 29.5 months (range: 3.0 to 42.0 months). All children were tested twice within two weeks. All children had either the BOS 2-30, the SON 2½-7, or the GOS 2½-4½ administered next to the BSID-II-NL. For 43 children, the developmental indices on the BSID-II-NL were compared with the developmental indices on the BOS 2-30 (Van der Meulen & Smrkovsky, 1983), the predecessor of the BSID-II-NL. The BOS 2-30 is considered to be a valid and reliable test by the Committee On Test Affairs Netherlands (COTAN) of the Dutch Professional Association of Psychologists, the NIP, 2001) but with antiquated standardisation. There is, of course, a strong similarity between the BSID-II-NL and the BOS 2-30 as they consist of a substantial number of identical items. However, this similarity is limited by differences in the administration procedure and by the fact that some of the items of the BOS 2-30 have been left out and new items were added. Table 13 shows an overview of the standard scores and correlations.

Table 13	BSID-II-NL vs. BOS 2-30: correlations and characteristics of the standard scores.					
		Standard scores				
		BSID-II-NL		*BOS 2-30*		
	Correlations	*Mean*	*SD*	*Mean*	*SD*	*n*
Mental scale	0.57*	108.47	15.33	114.63	16.87	43
Motor scale	0.52*	103.17	17.41	112.57	20.82	35

* Correlations significant with an alpha of 0.01 (one-tailed).

The correlation coefficients confirm the expectation that the tests are correlated, but not highly. Clearly there are differences in the content and structure of the test, but also the young age of the children is expected to be an influencing factor on the correlation. The correlations mentioned have been based on scores that did not take into account the unreliability of the instruments. If we were to take this into account, we would have to apply a correction for attenuation (Guilford, 1965). By applying this correction it is possible to determine how high the correlation between two variables would be if it were possible to determine both variables with perfect reliability. This correction gives a correlation of 0.68 on the mental and 0.67 on the motor scale. Table 13 also shows a significantly higher mean score on both scales of the BOS 2-30 than on the BSID-II-NL, over 6 DI points on the mental scale and nearly 10 points on the motor scale. These results confirm the expectation that the antiquated norms of the BOS 2-30 increase the chance of overestimating the developmental level of a child, both on the mental and the motor scales.

The BSID-II-NL developmental indices were also compared with scores on subtests of the GOS 2½-4½ (Neutel, Van der Meulen & Lutje Spelberg, 1996). The GOS 2½-4½ is a Dutch translation and adaptation of the Kaufman ABC (Kaufman & Kaufman, 1983). At the time that this validity study was conducted, the GOS was still widely used by Dutch institutions to assess the cognitive abilities of a child. However, the test material is no longer available due to a difficult license transition. If the test were still being used and were to be re-published, the data from this study would still be relevant.

The GOS 2½-4½ received a good evaluation by COTAN (Evers, Van Vliet-Mulder & Groot 2000) on the most important parts (norm scores, reliability and construct validity). The subtests of the GOS are organised around two ways of processing information: simultaneous (GOS 2½-4½ sim.) and sequential (GOS 2½-4½ seq.). All subtests together form the cognitive scale (GOS 2½-4½ cog.). A significant connection between especially the sequential scale of the GOS and the motor scale of the BSID-II-NL is expected. The motor planning and coordination skills as they appear in the motor scale of the BSID-II-NL specifically target the ability to order information sequentially and then process it. In table 14 means and standard deviations of test results on the GOS and BSID-II-NL are listed and in table 15 the correlations between both tests are listed.

The high mean scores on the different subtests of the GOS in relation to the test results on the BSID-II-NL stand out. For the cognitive scale of the GOS, for instance, the mean standard score is 118.63, for the mental scale of the BSID-II-NL the mean is 110.29. The mean score on the GOS even exceeds the range of normal development (85-115).

Table 14 **BSID-II-NL vs. GOS 2½-4½: distribution of standard scores.**

	N	*Mean*	*Standard deviation*
GOS 2½-4½ sim.	28	118.71	13.99
GOS 2½-4½ seq.	24	112.29	18.42
GOS 2½-4½ cog.	24	118.63	13.82
BSID-II-NL MS	28	110.29	11.37
BSID-II-NL MR	28	100.89	14.92

MS = mental scale, MR = motor scale, sim. = simultaneous scale, seq.= sequential scale, cog. = cognitive scale.

Table 15 **BSID-II-NL vs. GOS 2½-4½: correlations.**

		BSID-II-NL	
		Mental scale	*Motor scale*
GOS 2½-4½	Sim.	-0.06	-0.13
GOS 2½-4½	Seq.	0.33	0.38*
GOS 2½-4½	Cog.	0.19	0.23

* Correlation significant with an alpha of 0.05 (one-tailed). Sim. = simultaneous scale, seq. = sequential scale, cog.=cognitive scale.

Table 16	BSID-II-NL vs. SON-R 2½-7: distribution of standard scores.		
	N	*Mean*	*Standard deviation*
SON 2½-7 (SON-IQ)	28	106.18	14.00
BSID-II-NL MS	28	105.29	17.01
BSID-II-NL MR	22*	105.86	7.12

* The motor scale was administered to only 22 of the 28 children. MS = mental scale, MR = motor scale.

Table 17	BSID-II-NL vs. SON-R 2½-7: correlations.	
	BSID-II-NL	
	Mental scale	*Motor scale*
SON 2½-7 SON-IQ	0.55*	0.56*

* Correlations significant with an alpha of 0.01 (one-tailed).

Table 15 shows that the correlation between the different subtests of the GOS and the mental and motor scales of the BSID-II-NL is unexpectedly low. Even the correlation between the motor scale of the BSID-II-NL and the sequential scale of the GOS is low, although significant. Between the mental scale of the BSID-II-NL and the cognitive scale of the GOS a very low correlation of 0.19 was found. Corrected for attenuation, the coefficients for the mental scale are: -0.07 (sim.), 0.41 (seq.) and 0.22 (cog.) and for the motor scale: -0.16 (sim.), 0.48 (seq.) and 0.28 (cog.).

A comparison was also made between the test results on the BSID-II-NL and the SON-R 2½-7 (Tellegen, Winkel, Wijnberg-Williams & Laros, 1998). The SON-R 2½-7 is an instrument for diagnostic testing of children in the age group 2½ to 7 years, to be administered individually. COTAN (Evers, Van Vliet-Mulder & Groot, 2000) evaluates this test as 'good' on all parts. The achievements of a child on the six subtests can be summarised in an intelligence score, the SON-IQ. High correlation is expected, especially with the mental scale of the BSID-II-NL. Means and standard deviations of the developmental indices are listed in table 16. Correlations between the two tests are listed in table 17.

In calculating the SON-IQ, 'inflation correction' of the IQ was taken into account. The manual (Tellegen, Winkel, Wijnberg & Laros, 1998) states that the norm scores of the SON-R 2½-7 will age by 1 IQ point every three years.

The correlation found is sufficient. If we correct for the unreliability (correction of attenuation) of both tests the coefficients become 0.65 (mental scale) and 0.69 (motor scale).

Conclusion

The reliability of the BSID-II-NL was tested in a number of ways. The mean *internal consistency* coefficient of the mental scale is sufficient: 0.79 (range 0.51-0.96). The same holds for the motor scale: 0.74 (range 0.26-0.88). The lowest coefficients are found in the youngest age groups. This means that especially in those groups, the tests provide a less accurate picture of the general development of the child in comparison with older age groups. Care should therefore be taken when interpreting the data. Especially for children up to 12 months of age on the mental scale and the first few months on the motor scale, the achieved scores can be significantly higher or lower on re-testing.

The *test-retest correlations* are reasonable for the mental scale (0.75) and high for the motor scale (0.80). The results of the second test administration are higher on the mental scale in 73% of all cases and in 58% of the cases on the motor scale. This indicates that in general, experience from earlier testing has a positive influence on the scores of the child. For young children, this might be because of increased familiarity with the testing situation, and in the case of older children by increased familiarity with the test items. For the behaviour rating scale, test-retest correlations are low (0.49) for the factor score 'adaptation to the environment' and for the total score (0.67). The low correlations correspond to expectations. The items of the behaviour rating scale are not scored dichotomously but on a scale of 1 to 5, leaving more room for interpretation, and items are less accurately described than in the mental and motor scales. The factor 'quality of motor skills' can apparently be determined more objectively, correlation for this factor is high (0.83).

Our *stability* study showed a positive outcome. The prediction value of the mental scale at 18 months for the results at 30 months is high (0.58) when taking into account the young age of the children at the time of the first test and the long time interval (12 months) between the first and second test. For the motor scale it is low (0.33). This indicates that the skills measured by the mental scale at 18 months have a higher prediction value for the more complex skills at 30 months than the skills measured by the motor scale. The results on the motor scale at 18 months have to be viewed as a random indication with a low prediction value for future development, even more so than results on the mental scale.

For the mental scale the *inter-rater reliability* shows that the test results of a child are hardly influenced by the person who administers the test: correlation between two test administrations by different examiners is high (0.81). This also holds, albeit a little less pronounced, for the motor scale. Correlation here is reasonable (0.77).

Construct validity for the BSID-II-NL was tested by comparing the scores on the BSID-II-NL with scores on similar tests inside of two weeks. From the comparison of scores between the BSID-II-NL and the BOS 2-30 it is apparent that the tests are substantially related to each other ($r = 0.57$ on the mental scale and $r = 0.52$ on the motor scale), but less than should be expected based on the similarities between the tests. Apparently, the structure and content of the test has been changed to such an extent that the BSID-II-NL has become a similar test with the same background, but not simply a successor to the BOS 2-30. A comparison to the GOS 2½-4½ shows that the tests have little in common, despite the claim of both tests that they examine cognitive development. Only the motor scale of the BSID-II-NL and the sequential scale of the GOS have significant correlation. Between the mental scale and the total cognitive score of the GOS, there is a correlation of only 0.19. These results suggest that there is a little overlap in the content of the two tests. A possible explanation might be that the GOS measures cognitive development in a different way to the BSID-II-NL. The GOS specifically tests the child's ability to process information, whereas the BSID-II-NL tests more general skills. When taking into account that the SON-R 2½-7 differs significantly from the BSID-II-NL because of its specific non-verbal character and that, contrary to the BSID-II-NL, the SON-R 2½-7 requires an adaptive test procedure, the correlation between scores on both tests is more than reasonable. Correlation on the mental scale is 0.55 and on the motor scale 0.56. Higher correlations were not expected, because the subtests of the SON-R 2½-7 are more directed at solving problems that require spatial comprehension and the ability to reason abstractly and concretely. These skills are also tested in the BSID-II-NL, but to a lesser degree. Contrary to the SON-R 2½-7, language skills form an important part of the test, especially for children over the age of 2½ years.

There is a non-verbal version of the mental scale of the BSID-II-NL. This non-verbal version comprises a selection of, in some cases adapted, items of only the mental scale. The non-verbal version has norms for the age range of 12 to 30 months. A comparison of non-verbal developmental indices between the BSID-II-NL and the SON-IQ was not possible because of non-overlapping age ranges of both tests (the SON can only be administered to children aged at least 30 months). Mean scores on the tests used for validity determination were higher than the scores achieved on the BSID-II-NL in every case. For the SON 2½-7 and the GOS 2½-4½, this confirms the general impression from practical experience that the scores on these tests are often somewhat 'flattering'. For the BOS 2-30 it indicates that its standardisation is in fact outdated. The high mean developmental indices on the BOS 2-30 confirm the upward drift of norm scores. This pattern, sometimes referred to as the Flynn effect (Flynn, 1999), has been demonstrated in several other cognitive tests for young children (Black & Matula, 2000). The drift is estimated at 3 IQ points per decade (Flynn, 1984, 1987).

Relation with background variables

The relevant background data are sex of the child, education level of the parents, the examiner and regional origin of the child. To determine the background data, parents filled in a questionnaire before the test was administered.

Sex

Table 18 shows mean standard scores of girls and boys on the mental and motor scales. The mean standard score on the mental scale over all age groups for girls is a little more than three points higher than the mean score for boys. On the motor scale girls on average also perform a little better than boys. The mean score is about 1.5 index points higher. Research shows that girls on average experience a somewhat faster development of language skills than boys (see for instance Verhulst-Schlichting, Morelli-Kayser & Peddemors-Boon, 1987). The large contribution of language development to general cognitive development might be a partial explanation for the difference between girls and boys on the

Table 18	Mean standard scores of girls and boys on the mental and motor scales.					
	Mental scale			*Motor scale*		
Sex	*n**	*Mean*	*SD*	*N*	*Mean*	*SD*
Girls	943	101.75	14.75	943	100.87	14.78
Boys	959	98.44	14.55	954	99.11	14.64

* Total number of tests does not add up to 1909 (the total number of participating children) because the sex is not known for some children.

Table 19	Results of a two-way variance analysis of the effects of sex and the interaction effects of sex and age on the test results for the mental and motor scales.						
		Mental scale			*Motor scale*		
Source	*df*	*F test*	*Sig.*	*Effect size (f)*	*F test*	*Sig.*	*Effect size (f)*
Sex	1	14.55	0.00	0.10	1.19	0.28	0.00
Sex * Age	16	2.01	0.01	0.14	1.10	0.35	0.10

df = degrees of freedom, Sig .= significance (p value), effect size f = 0.10 = small effect; 0.25 medium effect; 0.40 large effect.

mental scale. A two-way variance analysis over all age groups (table 19) shows that the influence of sex on test scores is only significant ($p < 0.01$) for the mental scale. The effect size (Cohen, 1988), however, is small: $f = 0.10$. For the interaction effect between sex and age the same holds: the influence is significant ($p < 0.05$), but the effect size for this effect is small: $f = 0.14$.

Regional origin

For the standardisation study for the BSID-II-NL it was assumed that regional origin of a child has no significant influence on test scores. Several studies, among others the standardisation of the BOS 2-30 and both SON tests, have not shown any significant differences between children living in different regions. In table 20, the children of the UMCG and the UoG were considered separate groups. The children tested in the UMCG usually came from the region around Groningen, while most of the children tested in the UoG study lived in the city itself.

The differences in mean scores on the mental and motor scales are significant on a 1% level (mental scale: $F\{3,1895\} = 4.72$; $p < 0.01$; motor scale $F\{3,1895\} = 10.85$; $p < 0.01$) But the effect of origin is small on both scales.

Table 20	Results of a one-way variance analysis on the effect of regional origin on the test scores on the mental and motor scales.						
		Mental scale			*Motor scale*		
Source	*df*	*F test*	*Sig.*	*Effect size (f)*	*F test*	*Sig.*	*Effect size (f)*
Origin	3	4.72	0.00	0.10	10.85	0.00	0.14

df = degrees of freedom, Sig. = significance, Effect size f = 0.10 = small effect; 0.25 medium effect; 0.40 large effect.

Table 21 **Results of a three-way variance analysis on the effect of the examiner on the results for the mental and motor scales (main effect and interaction-effect with age and sex).**

		Mental scale			*Motor scale*		
Source	*df*	*F test*	*Sig.*	*Effect size (f)*	*F*	*Sig.*	*Effect size (f)*
Examiner	18	3.70	0.00	0.20	4.97	0.00	0.25
Examiner * age	97	1.25	0.06		1.52	0.00	0.31
Examiner * sex	18	0.88	0.60		0.65	0.85	

df = degrees of freedom, Sig. = significance; Effect size f = 0.10 = small effect; 0.25 medium effect; 0.40 large effect.

Examiner effect

In total, 37 examiners participated in the project, divided over four study centres. To study the effect of the examiner on the score, a selection was made of examiners, based on number of tests they administered. A minimum number of 25 tests per examiner were used to reduce the effect of sample fluctuations.

Table 21 lists the results of a three-way variance analysis with examiner as main effect and age and sex as interaction factors. At a significance level of 1%, both for the mental and motor scales, the factor examiner has significant correlation with the mean developmental index. The effect size ranges from small (mental scale) to medium (motor scale). The interaction effect of examiner and age is only significant for the motor scale ($p < 0.01$). The effect size of this is medium. No statistically significant interaction effect was found for either scale for examiner and sex.

Education of parents

In table 5, the education level of the parents is listed in percentages. The education level of the parents was known for all children in the standardisation study, but these levels could only be linked to individual children for the UoG and UMCG groups. For this reason, the analysis of this background variable only uses data from these two study centres.

There are differences in the mean developmental indices when these are linked to the level of education of the father and the mother. However, these differences are not significant, as is shown in the results of the variance analysis listed in table 22.

Conclusion

Analysis of the influence of the background variables (sex, regional origin, examiner and the education level of the parents) shows that while in some cases a significant correlation was found, the effect size of this correlation was small. A medium interaction effect was only found between examiner and age on the test score for the motor scale. The effect of the examiner on the end result is probably caused by the fact that some examiners tested a large number of children in a certain age group. In such a case, experience may start to play a part and cause a difference between test results for examiners with more experience with certain age groups as opposed to more 'all-round' examiners. It should be noted that this effect cannot be attributed to the examiner alone, since the children tested may actually have differed in mental and motor

Table 22 **Results of a three-way variance analysis on the effects of the level of education of the parents on the test scores for the mental and motor scales.**

		Mental scale		*Motor scale*	
Source	*df*	*F test*	*Sig.*	*F test*	*Sig.*
Mother's education level	2	1.01	0.39	0.67	0.57
Father's education level	2	1.97	0.12	2.34	0.07
Education mother * father	4	1.74	0.08	1.36	0.21

df = degrees of freedom, Sig.= significance.

ability. The sufficient (motor scale) to high (mental scale) inter-rater reliability seems to confirm this.

Discussion

We have described the construction of the BSID-II-NL, the Dutch standardisation study, supplementary studies into the psychometric characteristics and the relation between the instrument and some background variables. The aim is to answer the research question: *Is the Dutch translation and adaptation of the BSID-II a valid and reliable instrument for individual developmental assessment?*

An extensive Dutch standardisation and results from psychometric studies support the qualification of the BSID-II-NL as a valid and reliable instrument to determine the general development of young children in the Netherlands. The main research question can be answered positively. The research is, however, subject to some limitations and discussion on the possibilities and limitations for its use. This article ends with recommendations for further study.

The standardisation study

By developing a Dutch standardisation, it has become possible to compare the test results of a Dutch child on the BSID-II with a relevant norm group. Contrary to many other European countries, it is customary (and imposed by the government) in the Netherlands to only use tests that have Dutch norms and have been evaluated with regards to validity and reliability of the Dutch version.

From a comparative analysis between test scores of Dutch children in the standardisation study and American children (scores were taken from the American manual of the BSID-II), the same image arose that was suggested by the pilot study. On average and especially in the first 18 months, there is a developmental delay in Dutch children when compared with American children. See appendix 3 for illustrations. Using American norms on Dutch children younger than about 30 months would lead to an underestimation of both their cognitive and motor levels in many cases. Although the difference was about the same for all age groups and a correction procedure would have been an option, it was decided to develop a Dutch standardisation. In the first place because, as described above, a test is not deemed fit for use in the Netherlands unless it has a Dutch standardisation and secondly because any change in the test procedure and item instruction is reason for performing a standardisation study on the adapted test materials.

In most intelligence tests, norms are calculated per age group. The calculated norm is then correct for the middle of the age group. The level, however, of children at the starting age of the age group is underestimated and the level of children at the end of the age group is overestimated (Tellegen, 2004). The BSID-II-NL improves accuracy of norms by listing norms per half month for children under 12 months of age and per month for children aged between 12 and 42 months. The standardisation method used further allows for correction of deviations caused by using age intervals, by determining standard scores based on the exact age. It is practically impossible to determine norms based on the age in days of a child from a printed norm table, but this problem can be solved by using computerised scoring software. The scoring software for the BSID-II-NL allows the calculation of developmental indices and the accompanying reliability intervals based on the age in days. The software offers simple transformation of raw scores into standardised scores, takes the rapid development of young children into account and solves the problem of big jumps in the successive norm table for, in particular, children younger than 12 months. Those big jumps can cause a large difference in the norm score for an age difference of one day. For instance, for a child aged 5 months and 23 days, the developmental index is read from the age group of 6 months. If the child has, for instance, a raw score of 55, this gives a developmental index of 88. If the child had been tested a day earlier, at 5 months and 22 days, the developmental index would have been read from the age group of 5.5 months. This would have given a developmental index of 106. Using the software, a score of 55 gives a developmental index of 97 for a child aged 5 months and 22 days and 94 for a child aged 5 months and 22 days.

Instrumental utility

The BSID-II manual presents high reliability and validity figures for the mental and motor scale, as well as the behaviour record. In each category, the average American reliability and validity coefficients are higher than the Dutch results. Despite the limited sample sizes, the Dutch results are sufficient to high in every category. From these results we conclude that the test is sufficiently reliable, but for some parts, low reliability should be taken into account in the interpretation of the results. This depends on the age of the child and the scales that were administered. Until an age of about 12 months, one should be careful in interpreting the score, especially on the mental scale, but also on the motor scale. The broad reliability intervals should be observed. The reliability of the behaviour rating scale was only partially examined. The test-retest reliability is high for the factor *motor quality,* but low for the other factors. Because of the content of the items and the way in which they are scored, a larger measure of subjectivity is expected in this scale than in the mental and motor scales. Follow-up studies are needed to confirm this. To assess construct validity, we examined the relationship between the BSID-II-NL and comparable established Dutch instruments. Because of the BSID-II-NL's unique ability to assess the developmental level of very young children, no other test is available except the BOS 2-30 until the age of 30 months. Comparisons can only be made for 30 months and older with instruments such as the SON-R 2½-7 and the GOS 2½-4½. The obvious comparison with the early child version of the Wechsler scales, the WPPSI, can not be made because it does not have a good Dutch standardisation and the Dutch version can only be used from the age of 4 years. Because of the quality of the test and its frequent use, most value is attached to the more than sufficient correlation between the BSID-II-NL and the SON-R 2½-7.

Use of the BSID-II-NL

The BSID-II-NL offers the opportunity to determine norm scores for cognitive and motor development in a standardised and (because of the attractive play materials) child-friendly way. In combination with the judgement of the behaviour of the child, these can be used in judging the way a child functions, compared with a relevant reference group (see also Neisworth & Bagnato, 1996). Dunst (1998) adds that the test is very suited for this purpose, but has limited use in providing practical advice. As well as its low prediction value, because of its broad set-up (it only distinguishes between the cognitive and motor domain), the test is primarily meant to determine the *current* level of development and the planning of intervention goals for the short term (see Black & Matula, 2000; Lichtenberger, 2005). The Bayley Scales are, however, commonly used to assess results (cognitive function) of high-risk infants for predictive purposes, despite the assertion that prediction of future intelligence is not the purpose of the test. Niccols and Latchman (2002) concluded from their research that BSID-II assessments of high-risk infants in the first year of life should not be used for predictive purposes. From about 24 months of age, the correlation between test results and subsequent preschool scores is higher. For instance Molfese and Keogh (1997) show considerable correlation between test results on the BSID-II and the Stanford-Binet IV after two years of age. Results, however, are ambiguous. When correlating BSID-II test results on 20 months of age with KABC test results at eight years of age in extremely low birth weight children, poor predictive validity was reported (Hack, et al., 2005). Our stability study, however, showed a positive outcome. The prediction value of the mental scale at age 18 months for the results at 30 months is high (0.58) when taking into account the young age of the children at the time of the first test and the long time (12 months) between the first and second test. For the motor scale it is low (0.33). This indicates that the skills measured by the mental scale at 18 months of age have a higher predictive value for the more complex skills at 30 months than the skills measured by the motor scale. The results on the motor scale at 18 months have to be viewed as a random indication with a low prediction value for future development, even more so than results on the mental scale.

In the Dutch version, for the mental and motor scale, the item difficulty is listed for each item. The item difficulty is the age (in months and weeks) at which 50% of the children in the standardisation study have the tested ability and therefore score positive on the item. This information offers the possibility to analyse test results for each item as well as for the entire test. Especially in the case of children for whom it is difficult to determine a standard score because the test could not be completely administered, it is desirable to be able to form an (albeit not very precise or valid) impression of the level of the child with regards to specific skills that are tested in the items that were administered. This allows a more differentiated diagnosis to be made and therefore increases the usability of the BSID-II-NL.

When using item difficulties in interpreting test results, one should remember that the item order is based on the item difficulties as determined in the American standardisation study. The Dutch item difficulties differ from the American ones. Because of this, the Dutch scoring form shows a sub-optimal order of degree of difficulty. Since the test procedure requires the administration of a restrictive set of items per age group, it was impossible to change the item order afterwards. A change in the composition of the age groups would necessitate a new standardisation study.

An important advantage of standardised tests is that each child is confronted with the same test conditions, the same materials and the same instructions. This increases the validity and reliability of the instrument and allows a fair comparison between the test score of a child and the norm scores of a relevant reference group. A disadvantage of a high level of standardisation, as in the case of the BSID-II-NL, is that it places high demands on the user. The division of items in age groups forces the user to repeatedly learn new groups of items when testing children of different ages, and to remember a lot of information. Because of this, it takes some time and a sizeable amount of experience to be able to administer the test smoothly and without having to rely on the manual pending the administration. It is strongly advised to take plenty of time to familiarise oneself with the test and to use the item instruction as a control when scoring the items. Videotaping test administrations can also be a big help.

Another disadvantage of standardisation is that 'examiners cannot accommodate the potentially different needs of children with disabilities' (Kelly-Vance, Needelman, Troia & Oliver Ryalls, 1999). For purposes of maintaining standardisation and the use of the norm tables, examiners cannot deviate from the administration procedures dictated in the test manual. The tests may be biased, and therefore not suited for children with communication or physical impairments as they require use of language and motor skills to perform (Linder, 1993; McCormick, 1996). When test results of a child with a developmental disorder need to be interpreted, one should take into account the specific changes needed to compensate for the limitations of the child. For instance, a child with hearing problems can greatly benefit from exaggerated gestures and the use of pantomime instructions and a child with bad eye-hand coordination will be able to show more when enlarged play materials are used that require less motor planning and accuracy. Notwithstanding these adaptations, Lichtenberger (2005, p. 205) states that still 'much information can be gleaned from traditional cognitive measures (...) but clinicians must use their judgements to determine whether standardised norms are applicable and valid'.

The BSID-II-NL in its standard form is unsuited for determining the developmental level of many groups of children with specific limitations. For instance, for children of 24 months and older, the test places high demands on both language understanding and language production, and many of the cognitive items include a large (fine) motor component. To meet the demand for standardised instruments suitable for testing children with specific limitations, adapted versions of the BSID-II-NL were constructed. By using adapted procedures, item instructions and play materials, it is possible to compensate as much as possible for the problems the child experiences in executing skills because of its limitations. These adaptations are developed for children who were born prematurely, children with hearing impairments and/or speech/language impediments (BSID-II-NL, non verbal), children with visual impairments (BSID-II-NL, Low Vision) and children with motor impairments (BSID-II-NL, Low Motor).

Recommendations

To increase the usability of the BSID-II-NL, both scientifically and clinically, it is advisable to further investigate the possibilities of determining scores in a more differentiated way on the mental and motor scales by using factor analysis. For instance, it is expected that for children older than 24 months, discernable developmental areas can be determined. To form more detailed theories on the development of young children and to increase practical usability, it is important to achieve a more differentiated insight into the course of the development of a young child. Also, to support the results described in this article, further studies into the validity and reliability of the BSID-II-NL should be carried out. This should not only entail using larger numbers of children in the samples, but also a more extended selection of tests for investigating validity and reliability, for instance tests with more limited goals, such as the Reynell test for language understanding and the Schlichting test for language production.

The sample would have been more representative if the ethnicity of the child had been included in the study. The fact that ethnicity was not included is not viewed as a problem with regards to possible bias in the understanding of item instructions or items specifically connected with language, since all children in the sample were learning Dutch as their first language. Additional research is needed to determine possible differences between children with Dutch and other ethnic backgrounds.

The BSID-III was published in the US in 2005. This replaces the older norms of the BSID-II (1993) and accommodates US regulations for determining the development of different skills: cognition, language, motor, social-emotional, and adaptive behaviours. An optional behaviour rating scale is also included. A general cognition score, as in the BSID-II, can no longer be determined, because language skills have been moved to a separate scale. The BSID-III consists of five scales with items increasing in difficulty. The age groups and therefore the basal and ceiling rules of the BSID-II have been replaced by starting instructions based on the chronological age of the child (instead of age group) and the subsequent determination of the basal level by the number of consecutive positive scores and the ceiling by the number of consecutive negative scores of a child.

The publication of the BSID-III does not mean that the BSID-II-NL is already outdated. The Dutch standardisation and norms were constructed recently and should therefore be valid for approximately ten more years. The possible publication of a BSID-III-NL will depend greatly on the possibilities the publisher offers with regards to a translation and conducting a large-scale standardisation study. The recent publication of the BSID-II-NL gives us the opportunity to wait a number of years for further international studies into and practical experience with the BSID-III before deciding whether or not to develop a Dutch version.

References

Bayley, N. (1949). Consistency and variability in the growth of intelligence from birth to eighteen years. *Journal of Genetic Psychology*, 75, 165-196.

Bayley, N. (1993). *Manual for the Bayley Scales of Infant Development – Second Edition*. San Antonio, TX: The Psychological Corporation.

Black, M. & Matula, K. (2000). *Essentials of Bayley Scales of Infant Development-II Assessment*. New York: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences. Second edition*. New Jersey: Lawrence Erlbaum Associates Publishers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Drenth, P. J. D. & Sijtsma, K. (1990). *Testtheorie: inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten: Bohn Stafleu Van Loghum.

Dunst, C. (1998). Review of the Bayley Scales of Infant Development-Second Edition. In J.C. Impare, & S. Plake (Eds.), *The thirteenth mental measurements yearbook* (pp. 92-93). Lincoln, NE: the University of Nebraska-Lincoln.

Enquête beroepsbevolking (1999). Rijswijk: Centraal Bureau voor de Statistiek.

Evers, A. van Vliet-Mulder, J.C. & Groot, C.J. (2000). *Documentatie van Tests en Testresearch in Nederland. Deel II.* Assen: Van Gorcum , & Comp. B.V. Amsterdam: Nederlands Instituut van Psychologen Dienstencentrum B.V.

Flynn, J. R. (1984). The Mean IQ of Americans: Massive Gains 1932-1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 2*, 171-191.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54,* 5-20.

Gauthier, S. M., Bauer, C. R., Messinger, D. S. & Closius, J. M. (1999). The Bayley Scales of Infant Development II: Where to start. *Journal of Developmental and Behavioural Pediatrics, 20*, 75-79.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education. Fourth Edition*. New York: McGraw-Hill Book Company, p. 486.

Hack, M., Taylor, G. H., Drotar, D., Schluchter, M., Cartar, L., Wilson-Costello, D., Klein, N., Friedman, H., Mercuri-Minich, N. & Morrow, M. (2005). Poor Predictive Value of the Bayley Scales of Infant Development for Cognitive Function of Extremely Low Birth Weight Children at School Age. *Pediatrics, 116*, 333-341.

Kaufman, A. S. & Kaufman, N. L. (1983). *K-ABC. Kaufman Assessment Battery for Children. Interpretative Manual.* Circle Prins, MN: American Guidance Service.

Kelly-Vance, L., Needelman, H., Troia, K. & Oliver Ryalls, B. (1999). Early Childhood Assessment: A Comparison of the Bayley Scales of Infant Development and Play-Basis Assessment in Two-Year Old At-Risk Children. *Developmental Disabilities. Bulletin, 27, 1-15.*

Laros, J. A. & Tellegen, P. J. (1991). *Construction and validation of the SON-R 5.5-17*. Groningen: Wolters-Noordhoff.

Lichtenberger, E. O. (2005). General Measures of Cognition for the Preschool Child. *Mental Retardation and developmental disabilities. Research Reviews, 11,* 197-208.

Linder, T.W. (1993). *Transdisciplinary Play-Based Assessment*. (2nd ed.) Baltimore: Paul H. Brookes.

Lienert, G. A. (1961). *Testafbau und Testanalyse*. Weinheim: Beltz.

McCormick, K. (1996). Assessing cognitive development. In M. McLean, D. B. Bailey, Jr. & M. Wolery (Eds.), *Assessing infants and preschoolers with special needs.* Englewood Cliffs, New Jersey: Merrill, an imprint of Prentice Hall.

Matula, K. & Aylward, G. (1997). Response to commentary: BSID-II. *Journal of Developmental Behavioural Pediatrics, 18*, 112-113.

Molfese, V. J. & Acheson, S. (1997). Infant and Preschool Mental and Verbal Abilities: How Are Infant Scores Related to Preschool Scores? *International Journal of Behavioural Development, 20,* 595-607.

Neisworth, J. T. & Bagnato, S. J. (1996). Assessment for early intervention: Emerging themes. In S. Odom , & M. McLean (Eds.), *Early intervention/early childhood special education: Recommended practices* (pp. 23-57). Austin, TX: PROED, Inc.

Neutel, R. J., van der Meulen, B. F & Spelberg, L. H. C. (1996). *Groningse ontwikkelingsschalen: handleiding, GOS 21/241/2*. Lisse: Swets Test Services (STS).

Niccols, A. & Latchman, A. (2002). The Stability of the Bayley Mental Scale of Infant Development with High Risk Infants. *The British Journal of Developmental Disabilities, 48,* 3-13.

NIP (2004). *Algemene Standaard Testgebruik NIP (AST-NIP).* Voorstel. Amsterdam: NIP.

Nunnally, J. C. (1978). Psychometric Theory. New York, London: McGraw-Hill.

Sarneel, N. *De nieuwe Bayley Ontwikkelingsschalen.* Unpublished master's thesis, University of Groningen, Groningen, the Netherlands.

Siegel, L. (1981). Infant tests as predictors of cognitive and language development at two years. *Child Development, 52*, 545-557.

Snijders, J.Th, Tellegen, P. J. & Laros, J. A. (1988). *Snijders-Oomen niet-verbale intelligentietest SON-R 51/2-17. Verantwoording en handleiding*. Groningen: Wolters-Noordhoff.

Ten Berge, J. M. F. & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43,* 575-579.

Tellegen, P. J., Winkel, M., Wijnberg-Williams, B. J. & Laros, J. A. (1998). *Snijders-Oomen Niet-verbale In-*

telligentietest, SON-R 21/2-7. Verantwoording en handleiding. Lisse: Swets , & Zeitlinger.

Tellegen, P.J. (2004). *De waan van het IQ*. Internet: www.testresearch.nl.

Van Eldik, M.C.M. (1998). *Meten van taalbegrip en taalproductie: constructie, normering en validering van de Reynell test en de Schlichting test voor taalproductie*. Groningen: Stichting Kinderstudies.

Van der Meulen, B.F. & Smrkovsky, M. (1983). *Bayley Ontwikkelingsschalen BOS 2-30*. Lisse: Swets , & Zeitlinger.

Van der Meulen, B.F., S.A.J., lutje Spelberg, H.C.L. & Smrkovsky, M. (2002). *Bayley Scales of Infant Development Second Edition - Nederlandse versie*. Lisse: Swets Testpublishers.

Verhulst-Schlichting, J.E.P.T., Morelli-Kayser, M. & Peddemors-Boon, M. (1987). Sex and family background in early language acquisition. In Brouwer, D. & Haan, D. (ed) *Women's language, socialization and self-image*. Dordrecht: Floris.

Washington, K., Scott, D.T., Johnson, K.A., Wendel, S. & Hay, A.E. (1998). The Bayley Scales of Infant Development-II and Children with Developmental Delays: A clinical perspective. *Developmental and Behavioural Pediatrics, 19*, 346-349.

Zwart, F. (2004). *Instrumentele utiliteit van de BSID-II-NL*. Unpublished master's thesis, University of Groningen, Groningen, the Netherlands.

Appendix 1 Examples of items on the three scales

Mental scale

168. Complete the pattern (seated)

Administration: Place the pegboard on the table, directly in front of the child.

Trial 1. Hold the four red pegs and the three blue pegs in your hand and say to the child:

I am going to use these coloured pegs to make a pattern.

Then place the pegs appropriately while saying:

First I put in a red peg and then a blue peg. Next I put in another red peg and then a blue peg. Now I put in a red peg.

Lay the remaining pegs in front of the child and say:

Which coloured peg comes next? You put the next peg in the hole.

Regardless of the child's response, present trial 2.

Trial 2. Hold three of the red pegs, the three blue pegs, and the two yellow pegs in your hand and say to the child:

Let's make another pattern.

Place the pegs appropriately while saying:

First I put in a red peg and then a blue peg and then a yellow peg. Next I put in another red peg and then a blue peg.

Lay the remaining pegs in front of the child and say:

Which coloured peg comes next? You put the next peg in the hole.

Recording form: Place a check mark in the space for trial 1 if the child puts in the blue peg; place a check mark in the space for trial 2 if the child puts in the yellow peg.
Scoring: Give credit if the child put in the blue peg in trial 1 and the yellow peg in trial 2.
Previous item in series: 162
Material: Pegboard, four red pegs, three blue pegs, and two yellow pegs.

Motor scale

87. Lace three beads (seated)

Caution: Take care to prevent the child from placing beads in its mouth.
Administration: Knot one end of each shoe string and place the beads on the table. Lace two beads on your string. Then give the child the other shoe string and three of the beads. Say to the child:

Here is a string for you. Put the beads on the string. Put them all on.

If the child places all three beads on the string, push the remaining three beads to the child and say:

Put these on. Put them all on.

If the child stops or takes off beads before placing three beads on the string, say:

Put some more beads on. See how many you can put on.

Recording form: Record the number of beads that the child laces.
Scoring: Give credit if the child puts at least three beads on the string at any one time.
Material: Two shoestrings and eight square beads.

Behaviour rating scale

13. Exploration of objects and/or surroundings

1-42 months: The degree to which the child actively seeks out new aspects of objects or the environment, including the child's visual, auditory and tactile exploration.
1 No exploration
2 One or two instances of exploration
3 Moderate exploration
4 Much exploration
5 Constant exploration

Appendix 2 Regression equations

Regression equations where:
x = raw score / 100
L = age in months / 12
DI = developmental index with a range of 55-145

Mental scale
Age 1 to 13 months
$DI = 89.726728 + 593.625717x + 394.439760x^4L – 653.997903L – 833.657562xL + 776.520418L^2 – 52.750843xL^4 – 96.627180x^3L^5$

Age 7 to 42 months
$DI = -155.754594 + 598.60036Lx + 53.488128x^3 – 468.854580xL + 59.396208L^2 + 70.017802xL^2 – 0.81628L^5$

Motor scale
Age 1 to 42 months
$DI = 36.714690 + 73.376303x – 4.550059x^2 + 0.003247x^5 – 327.194629L – 0.000461x^5 + 99.133583L^2 – 0.877068xL^4 + 0.163023xL^5$

Appendix 3 Comparative analysis between test scores of Dutch and American children.