Research in Developmental Disabilities 71 (2017) 109–119

Contents lists available at ScienceDirect

Research in Developmental Disabilities

journal homepage: www.elsevier.com/locate/redevdis

Research Paper

# Difference or delay? A comparison of Bayley-III Cognition item scores of young children with and without developmental disabilities

Linda Visser[a,*], Carla Vlaskamp[b], Cornelius Emde[c], Selma A.J. Ruiter[d], Marieke E. Timmerman[c]

[a] German Institute for International Educational Research (DIPF) and Centre for Research on Individual Development and Adaptive Education for Children at Risk (IDeA), Schloßstraße 29, 60486, Frankfurt am Main, Germany
[b] University of Groningen, Faculty of Behavioural and Social Sciences, Department of Special Needs Education and Youth Care, Grote Rozenstraat 38, 9712 TJ Groningen, The Netherlands
[c] University of Groningen, Faculty of Behavioural and Social Sciences, Department of Psychometrics and Statistics, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands
[d] De Kinderacademie Groningen, Herestraat 106, 9711 LM Groningen, The Netherlands

ARTICLE INFO

*Keywords:*
Developmental disabilities
Young children
Cognitive development
Developmental assessment
Differential item functioning

ABSTRACT

The "difference or delay paradigm" focuses on the question of whether children with developmental disabilities (DD) develop in a way that is only delayed, compared to typically developing children, or also qualitatively different. The current study aimed to examine whether qualitative differences exist in cognitive development of young children with and without DD on the basis of item scores on the Dutch Bayley-III Cognition scale. Differential item functioning was identified for 15 of the 91 items. The presence of DD was related to a higher number of Guttman errors, hinting at more deviation in the order of skill development. An interaction between group (i.e., with or without DD) and developmental quotient appeared to predict the number of Guttman errors. DD was related to a higher number of Guttman errors for the whole range of developmental quotients; children with DD with a small developmental quotient had the highest number. Combined, the results mean that qualitative differences in development are not to be excluded, especially in cases of severe developmental disabilities. When using the Bayley-III in daily practice, the possibility needs to be taken into account that the instruments' assumption of a fixed order in skill development does not hold.

## What this paper adds?

Within the "difference or delay paradigm", "delay" means that persons with developmental disabilities (DD) develop cognitive skills in the same order as persons without DD, but at a slower rate and with a lower ceiling. "Difference" means that, in addition to the delay, there are qualitative differences in development, for example in the order in which skills develop. The discussion has not yet been solved, but differences can clearly not be excluded.

Standardised developmental assessment instruments assume consistency between children's order of skill development. If "difference" appears the case, this assumption does not hold, which yields problems for the tests' validity. The assumption, however, has

* Corresponding author.
*E-mail addresses:* Linda.Visser@dipf.de (L. Visser), C.Vlaskamp@rug.nl (C. Vlaskamp), c.emde@me.com (C. Emde), Selma.Ruiter@dekinderacademie.com (S.A.J. Ruiter), M.E.Timmerman@rug.nl (M.E. Timmerman).

http://dx.doi.org/10.1016/j.ridd.2017.09.022
Received 9 March 2017; Received in revised form 26 September 2017; Accepted 29 September 2017

never been studied for the Bayley-III, a high standard developmental assessment instrument that is widely used to assess children with DD. The current study aimed to do this by comparing the Cognition scores of young children with and without DD. We found Differential item functioning between the groups. The score patterns of children with DD with a low developmental quotient appeared to deviate the most from the expected order of skill development. In case of developmental quotients approaching the normal range, the deviation was still larger for children with DD than without DD. The results indicate that caution is needed when assessing children with DD with low developmental quotients with the Bayley-III: the tests' assumptions might not be valid, causing the risk of overestimating skills below the basal and missing skills above the ceiling.

## 1. Introduction

Is the development of children with developmental disabilities only delayed compared to children with a typical development, or also qualitatively different? This is the focus of the "difference or delay paradigm", which has originated in the literature in the late 1970s and still remains unsolved.

Weisz and Zigler (1979) have formulated the "similar sequence hypothesis", which states that persons with intellectual disability (ID) develop cognitive skills in the same order as persons without ID, but at a slower rate and with a lower ceiling. Their literature review included studies using Piagetian tasks with diverse target groups, including persons with profound ID. The results support the similar sequence hypothesis for persons with ID, both with and without organic causes.

A review including studies with information-processing tasks (Weiss, Weisz, & Bromfield, 1986) showed completely different results: persons with ID performed worse on the tasks than persons without ID matched on cognitive developmental level, especially in the higher level ranges. Performance was especially impaired for discrimination and memory tasks. This difference in performance in some tasks, but not others, hints at qualitative differences and does thus not support the similar sequence hypothesis.

The answer to the "difference or delay"-question could thus depend on the type of tasks studied. It could also depend on the age of the child, the specific domain studied, and the type of disorder underlying the ID (Hodapp & Burack, 1990). Development until infancy seems to be delayed, while later development seems to be qualitatively different. Biologically based domains tend to develop delayed, while those that are mainly influenced by the environment more often show qualitative differences (Hodapp & Burack, 1990). In terms of the type of disorder, specific genetic disorders are related to specific strengths and weaknesses (Dykens & Hodapp, 2001), which can lead to qualitative differences in development. Research results have shown qualitative differences in the developmental order and processes of children with Down syndrome (Hasan & Messer, 1997; Lauteslager, 2000; Morss, 1985; Wishart, 1993). Qualitative differences are also found outside the domain of ID and cognition, for example in the language (Pérez-Pereira and Conti-Ramsden, 1999) and motor development (Reimer, Cox, Boonstra, & Nijhuis-Van der Sanden, 2015) of children with blindness and premature birth (Van Braeckel et al., 2010).

Although many authors still conclude that the similar sequence hypothesis is mostly supported (Bennett-Gates & Zigler, 1998; Facon, 2008), especially in cases of familial as opposed to organically caused ID (Burack, Russo, Flores, Iarocci, & Zigler, 2012), research results hinting at qualitative differences in cognitive development are numerous. The sequence of development as well as the underlying processes can be different (Nabuzoka, 2008).

This raises questions regarding the use of standardised developmental assessment instruments for assessing children with ID. In daily practice, standardised developmental assessment instruments are used to diagnose developmental delay. Even though (criterion-referenced) instruments that are specifically developed for persons with ID exist (AAIDD, 2008; Buntinx & Schalock, 2010), standardised developmental assessment instruments are also applied to estimate the developmental level in children who have already been identified with ID or developmental disabilities.

In the current article, we focus on young children with Developmental Disabilities (DD), defined as "severe chronic disabilities that can be cognitive or physical or both" (AAIDD, 2017). ID thus falls under the umbrella term of DD (Schalock et al., 2010; AAIDD, 2017), which means that a person with ID always has DD, but a person with DD does not always have ID, although the overlap is large (AAIDD, 2017). The term DD is commonly used for children who are too young for more specific diagnoses to be identified. Due to the large overlap, children with DD are often diagnosed as having ID when they are older, which is usually around school age. The term DD is also used for describing disabilities which are broader than only intellectual, such as those including physical disabilities. A disability in one area (e.g., motor) can influence the development in other areas (e.g., cognitive) and possibly cause qualitative differences therein. This effect is strengthened by the large degree of interrelatedness of developmental areas in young children (Couturier & Tak, 2002), which is even more pronounced in cases of DD (Houwen, Visser, Van der Putten, & Vlaskamp, 2015). The construct of DD covers this complexity. The implication is that the possibility of qualitative differences in development needs to be taken into account in cases of DD, like in cases of ID.

The most widely used instrument for developmental assessment in young children with DD is the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III; Bayley, 2006). It can be used for assessing the development up to a (developmental) age of 3½ years. The instrument is based on the assumption that children develop skills in a fixed order (e.g., "similar sequence"): The test items are ordered on the basis of their difficulty, determined by scores of children in the standardization sample, most of whom have a typical development. Test procedures are based on the assumption that this order is the same for all children: basal and ceiling rules determine which items are administered. It has not yet been studied to what extend this assumption holds in cases of DD.

The Bayley-III has mainly been evaluated for use with children born preterm (Reuner, Fields, Wittke, Löpprich, & Pietz, 2013; Spencer-Smith, Spittle, Lee, Doyle, & Anderson, 2015; Velikos et al., 2015) or with medical conditions (Acton et al., 2011; Hallioglu et al., 2015; Komur et al., 2013). These studies, as well as the technical manual of the Bayley-III (Van Baar, Steenis, Verhoeven, Hessen, 2014), describe the scores of the children on the Bayley-III. However, as Burack and colleagues (Burack et al., 2012, p. 5)

note, information regarding lowered performance in children with disabilities "of course, is not at all surprising, and not at all informative". Instrument suitability for children with DD needs to be verified in a different way than by only describing their test scores.

Research on earlier versions of the Bayley scales shows that suitability is not self-evident. Wishart and Duffy (1990) studied the validity of the first version of the Bayley scales (Bayley, 1969) for children with Down's syndrome. Their most important finding was that there was a very large variation in scores on the item level among children who were tested twice. This points at little stability in performance of children with Down's syndrome, which has severe consequences for test validity. Moore, Goodwin and Oates (2008) have noted that the Cognition items of the second version (BSID-II; Bayley, 1993) pertain to different constructs, including social and motor-related skills. This can lead to inaccurate measures of cognition in children with Down syndrome, as the level of social and motor development often differs from the level of cognitive development. In the Bayley-III, Cognition items rely on social skills to a lesser extent, as separate Communication scales are introduced, but still rely on motor skills (Visser, Ruiter, Van der Meulen, Ruijssenaars, & Timmerman, 2013).

The aim of the current study is to examine whether qualitative differences exist in cognitive development of young children with and without DD, as measured by the Bayley-III Cognition items. In other words, we aim to study the assumption of the Bayley-III that children develop skills in a fixed order. The study results will thereby be indicative of the validity of the Bayley-III for children with DD and may add to our knowledge regarding the difference or delay paradigm. Because the "difference or delay paradigm" and "similar sequence hypothesis" focus primarily on cognitive development, we will focus on the Cognition scale of the Bayley-III.

The following research questions will be addressed:

1. Do children with and without DD differ in the extent to which their Bayley-III Cognition test scores are in agreement with the assumption of a fixed order of skill development?
2. If differences are found, are these related to particular kinds of cognitive skills that develop faster or slower in one of the groups?
3. If differences are found, can they be explained by variables like diagnosis or the degree of developmental delay?

The Bayley-III is an eclectic instrument, which means that it is based on a large number of developmental theories. The items do not only include Piagetian tasks, but also many tasks related to information processing (Van Baar, Steenis, Verhoeven, Hessen, 2014). Therefore, we hypothesise that the scores of children with DD are in agreement with the assumption of a fixed order of skill development to a lesser extent than the scores of children without DD. Regarding question 2, we expect more differences in tasks related to information processing (e.g., memory and discrimination tasks) than in tasks related to the theory of Piaget (e.g., object permanence). Regarding question 3, we expect more differences in children with DD related to organic causes (Burack et al., 2012), such as is the case in genetic syndromes, and in children with a lower degree developmental delay (Weiss et al., 1986).

## 2. Material and methods

### 2.1. Participants

The study included two groups of participants: children with DD (DD group) and children without any known DD (control group).

#### 2.1.1. DD group

We used the data of test administrations with the Dutch version of the Bayley-III (Bayley-III-NL) from the study into the Special Needs Addition (SNA; Visser, 2014) and the Dutch standardization study (Van Baar, Steenis, Verhoeven, Hessen, 2014). Children were included in these studies when they had diagnosed or suspected DD and an estimated developmental age between 0 and 3;6 years, as judged by the child psychologist or special needs specialist who referred the child. This means that children with a calendar age above 3;6 years were also included, which is in line with the use of the Bayley-III in daily practice as well as instructions in the manual (Van Baar, Steenis, Verhoeven, 2014).

From the SNA-study, we included test data of children who were tested with the standard ($n$ = 136), Low Verbal (LVe; in cases of a speech-/language impairment; $n$ = 66), Low Motor/Vision (LM/LVi; in cases of motor and/or visual impairment; $n$ = 31) or dynamic ($n$ = 58) version of the Bayley-III-NL. Data from tests with the LVe and LM/LVi versions could be included because research results have supported that they are parallel versions to which the standard norms apply (Visser et al., 2013; Visser, Ruiter, Van der Meulen, Ruijssenaars, & Timmerman, 2015). Data from tests with the dynamic version could be included, because it includes a standard Bayley-III assessment. The children in the LM/LVi-study were all tested with both the standard and LM/LVi-version. We only included the results on the more valid (Visser et al., 2013) LM/LVi-version, and only if this was administered as first, to prevent an influence of a learning effect.

From the standardization study, we included the test data of 46 children from the clinical groups, who had Down syndrome or extreme premature birth (< 32 weeks gestational age). In line with the instructions in the manual (Van Baar, Steenis, Verhoeven, 2014), the age correction for prematurity was applied until the calendar age of 24 months. We did not assess moderate premature birth (32–36 weeks gestational age) as a reason to include a child in the DD group, because only extreme prematurity is related to DD (Johnson, Marlow 2011). We excluded children of whom no information was available on diagnosis or disability.

In total, the DD group included 337 children (194 boys; 143 girls) from all over the Netherlands. The average age was 4;1 years ($SD$ = 2;1, range = 0;3–11;2). The ethnic background was: 295 Dutch, 11 non-western, 8 western foreign, 4 from the Netherlands Antilles or Suriname, and 19 unknown. The educational level of the mother was low ($n$ = 83), medium ($n$ = 113), high ($n$ = 89), or

**Table 1**
Frequency of diagnoses and impairments of the children in the DD group.

| | | n |
|---|---|---|
| Diagnose | None | 99 |
| | Pervasive developmental disorder | 29 |
| | ADHD | 4 |
| | Other clinically diagnosed disorder | 1 |
| | Hydrocephalus | 10 |
| | Cerebral palsy | 8 |
| | Down syndrome | 52 |
| | Angelman syndrome | 3 |
| | Phelan McDermid Syndrome | 23 |
| | Other genetic syndrome | 24 |
| | Other (not in list above; known with author) | 47 |
| | Unknown | 50 |
| Impairment | Motor | 81 |
| | Visual | 43 |
| | Speech-language or hearing | 142 |
| | None | 69 |
| | Unknown | 106 |
| Premature birth | No | 219 |
| | Moderate or late | 36 |
| | Extreme | 38 |
| | Unknown | 44 |

*Note.* The numbers do not add up to 337, because diagnoses and impairment can overlap. ADHD = Attention Deficit Hyperactivity Disorder.

unknown ($n$ = 52). Table 1 gives information about the diagnoses and disabilities of the children, as far as these were known at this young age.

Test data of 19 children who were tested a second time, were also included. To rate out learning effects on test results, the time period between these two test administrations was at least 6 months (Van Baar, Steenis, Verhoeven, 2014). The total number of test administrations thereby adds up to 356.

#### 2.1.2. Control group

The control group was formed on the basis of test administrations with the Bayley-III-NL from the Dutch standardization study (Van Baar, Steenis, Verhoeven, 2014), excluding the clinical group. It consists of 1633 children (827 boys; 806 girls) between 16 days and 3;6 years of age ($M$ = 1;5, $SD$ = 1;0) and without any known DD. The children in the control group could be compared to those in the DD group because they both had a cognitive developmental level in the mentioned age range.

The control group is representative for the Dutch population (Centraal Bureau voor de Statistiek, 2011; Van Baar, Steenis, Verhoeven, Hessen, 2014). Their ethnic background was: 1305 Dutch, 206 non-western, 74 western foreign, and 48 from the Netherlands Antilles or Suriname. The educational level of the mother was low ($n$ = 208), medium ($n$ = 600), or high ($n$ = 825).

### 2.2. Instrument

The Bayley-III-NL is a developmental assessment instrument with norms for children with a calendar age or expected developmental age up to 3;6 years. It contains five individually administered scales: Cognition, Receptive Communication, Expressive Communication, Fine Motor, and Gross Motor skills. The Communication and Motor scales were only administered when there was sufficient time. We only considered the Cognition scores.

The Cognition scale consists of 91 items scored dichotomously (positive = 1/negative = 0) and ordered on the basis of difficulty. Basal and ceiling rules determine which items are administered. Items before the first administered item are automatically scored 1; items after the last administered item are automatically scored 0. The raw score is determined by counting the number of items with a 1-score.

The Bayley-III-NL is identical to the original version from the United States (Bayley, 2006), except for the language used. It has been standardised on the basis of a sample of 1953 children. The correlations of test scores with these of comparable instruments (BSID-II; Van der Meulen, Ruiter, Lutje Spelberg, & Smrkovsky, 2002, and WPPSI-III-NL; Wechsler, 2009) appear moderate (0.30–0.50). The inter-item relations range from 0.53 to 0.98, with an average between 0.81 and 0.90 for the scales. Test-retest reliability coefficients range from 0.38 to 0.86 and increase with age (Van Baar, Steenis, Verhoeven, Hessen, 2014).

### 2.3. Procedure

In the original SNA-study, psychologists or special needs specialist of organisations supporting young children with DD referred children on the basis of the earlier mentioned inclusion criteria. The researchers chose the most suitable version of the Bayley-III-NL

per child on the basis of the referral information. The tests were administered by the referrer, a test assistant from the organisation, or an advanced university student who had received a training by the researchers. The tests took place at the organisation ($n$ = 281), the university ($n$ = 19), or the child's home ($n$ = 52). In four cases, the location was different or unknown.

For details about recruitment of participants and procedures of testing in the standardization study, we refer to the technical manual of the Bayley-III-NL (Van Baar, Steenis, Verhoeven, Hessen, 2014).

To develop the data files that form the basis for the current study, we removed the data from unreliable test administrations ($n$ = 6 for the DD group; $n$ = 29 for the control group), defined as having more than 3 non-scored items, in which the tester forgot items or the child did not cooperate. In the numbers of participants mentioned previously, these unreliable test administrations were already excluded.

To characterise the children in the DD group, we looked up the index scores using QGlobal (the online scoring platform of the Bayley-III-NL with day norms). We did not include information about the index scores of children in the control group. These would most likely be a bit higher than 100 because the children with special needs were excluded.

Due to missing information about pregnancy length, we were not able to correct the index score for prematurity, as is usually advised for children up to 24 months of age (Van Baar, Steenis, Verhoeven, 2014).

### 2.4. Analyses

Given the difference in sample size between the two groups, we have chosen analyses that are not vulnerable to such differences in sample size.

First, we calculated descriptive statistics of the calendar ages and raw scores of the children in both groups. We calculated the developmental quotient by dividing the age equivalent related to the raw Cognition score by the child's actual age and multiplying with 100. The developmental quotient is used in daily practice as well as research and has supported predictive validity (Milne, McDonald, & Comino, 2014). It gives an indication of the rate of development and forms the single available standardised test score if a child is older than 3;6 years.

The subsequent analyses are based on both the scores on items that were actually administrated and those that were imputed using the basal and ceiling rules, thereby adhering to the standard scoring rules of the Bayley-III. The proportion of items that were administered versus imputed needs to be taken into account while interpreting the results. Therefore, we mapped the number of times each of the 91 items in the Cognition scale had been administered.

To study differences between the two groups in the extent to which the test scores reflect a fixed order of skill development, we looked into Differential item functioning (DIF) and the number of Guttman errors.

DIF analysis is a statistical approach for evaluating the suitability of items of a scale for a specific group. DIF is said to occur when individuals from two subgroups, but with the same level on the latent trait (i.e., the trait being measured), have different probabilities of correctly responding to an item (Bechger, Maris, & Verstralen, 2010; Camilli & Shepard, 1994; Finch, Barton, & Meyer, 2009). Commonly, a distinction is made between uniform and nonuniform DIF, meaning that the degree of DIF is or is not, respectively, the same across different levels of the latent trait (Finch et al., 2009).

Studies into DIF have been used for evaluating the suitability of assessment instruments for specific groups. Examples are a study into a language test for infants with autism (Bruckner, Yoder, Stone, & Saylor, 2007), an accommodated test for students with special needs (Finch et al., 2009; Randall, Cheong, & Engelhard,George, Jr., 2011), a large print or braille version of a test for students with visual impairment (Stone, Cook, Cahalan-Laitusis, & Cline, 2010), and a play assessment for children with specific language impairment (Lautamo, Laakso, Aro, Ahonen, & Törmäkangas, 2011). Regarding the Bayley-III, a study was done in which DIF-analysis was used to compare an accommodated version for children with speech/language problems with the standard version (Low Verbal version; Visser et al., 2015).

Techniques for studying DIF are built upon the assumption that all items in a test measure one characteristic of the target population of the test, which is called the latent trait. In case of the Bayley-III, this assumption is questionable: do all test items measure one and the same characteristic? This is not self-evident, as they are meant to measure cognitive development, which is a construct that changes by definition. Given this uncertainty, a discovery of DIF in the Bayley-III could not only hint at unsuitability of the item for the target group. It could also hint at qualitative differences in development of the target group, compared to children with a typical development. Although DIF-analysis is typically used to study item characteristics, our idea is therefore that DIF-analysis of Bayley-III scores can also be used to study qualitative differences in developmental course: if a child develops in a way that is different from what is regarded normal, the response to a test item might be different as well.

We started the analyses by checking the assumption of unidimensionality of the Cognition scale among the control group using Automated Item Selection Procedure (AISP) for Mokken Scale Analysis in R (Van der Ark, 2012). For studying DIF, we used a combination of the Combined Decision Rule (CDR) and logistic regression. We used R (R Core Team, 2016) with the difR package (Magis, Beland, & Raiche, 2015). The CDR is based on a combination of the Mantel-Haenszel (MH) and Breslow-Day (BD) procedures, as MH is powerful for detecting uniform DIF and BD for detecting nonuniform DIF (Güler & Penfield, 2009). We used item purification of at least 100 iterations. In the Mantel-Haenszel test, we used the Yates correction. For calculating effect sizes, we used the ETS Delta scale (ETS Δ; Holland & Thayer, 1985), which is commonly used in this context (Holland & Thayer, 1988). For the exact formulas, we refer to Magis, Béland, Tuerlinckx, and De Boeck (2010).

Logistic regression is an alternative, also recommended method for detecting DIF in dichotomous items (Freeman & Miller, 2001). We applied logistic regression to investigate which items show DIF and thereby used those items not identified as containing DIF by the CDR as anchors. Subsequently, we studied the content of the items that were identified as containing DIF by the CDR and/or

**Table 2**
Descriptive statistics per group.

| | DD Group | | Control Group | |
|---|---|---|---|---|
| | *N* = 356 | | *N* = 1633 | |
| | *M* (*SD*) | range | *M* (*SD*) | range |
| Administered items (*n*) | 22 (8) | 4–48 | 23 (7) | 7–49 |
| Calendar age (y;m) | 4;2 (2;2) | 0;3–11;2 | 1;6 (1;0) | 0;1–3;7 |
| Developmental age (y;m) | 1;10 (0;10) | 0;1–3;7 | 1;6 (1;1) | 0;0–3;7 |
| Raw score Cognition | 58 (16) | 7–86 | 47 (24) | 2–87 |
| Developmental quotient | 54 (28) | 3–119 | 100 (20) | 7–253 |

*Note.* DD = Developmental Disabilities, *M* = Mean, *SD* = standard deviation, y;m = years; months, Developmental age = developmental age equivalent based on the raw Cognition score.

logistic regression to search for an explanation for the DIF.

The number of Guttman errors is "the number of item pairs with a 0 on the easier item and a 1 on the more difficult item" (Meijer, 1994, p. 311). It is a useful statistic for determining the suitability of items. Weisz and Zigler (1979) already proposed applying Guttman analysis to test the "similar sequence hypothesis". We have calculated the number of Guttman errors per child using the R package MSA (Van der Ark, 2012). To correct for differences in the number of administered items, we calculated the ratio of Guttman errors relative to the number of items administered.

To investigate which variables predict this Guttman error ratio, we performed a multiple regression analysis (using the function LM in R), with as main effect predictor the centred developmental quotient, and the dummy variables group (including the possible interaction with centred developmental quotient), prematurity, motor impairment, visual impairment, auditory or speech/language impairment, pervasive developmental disorder, Down's syndrome, Phelan McDermid syndrome, and other genetic syndrome. The developmental quotient was centred around the average in the population, which is 100. By including the developmental quotient, we were able to account for the difference between the two groups in terms of calendar age. We did not include the other diagnoses that are mentioned in Table 1 as predictors, because the number of observations was lower than the recommended 15–20 (Hair Black, Babin, Anderson, & Tatham, 2006). We also performed a multiple regression with only group and the centred developmental quotient as the predictors, and their interaction.

## 3. Results

### 3.1. Descriptive statistics

Table 2 shows the means, standard deviations, and ranges of the number of administered items, calendar ages, developmental age equivalents, raw scores, and developmental quotients, per group. The calendar ages and developmental quotients clearly differ between the groups, which is a logical consequence of the sample characteristics. The number of administered items does not differ between the groups ($t = -0.368$; $p = 0.713$). The developmental age equivalent is higher in the DD group than in the control group ($t = -6.658$; $p < 0.001$), which is also reflected in the different mean raw score.

In Fig. 1, the (estimated) distributions of the developmental quotients in the two groups are shown graphically. It shows that the children in the DD group are behind in their development, while the developmental quotients of the control group have an about normal distribution around the average of 100.

The number of times an item was actually administered or imputed according to the basal and ceiling rules differed per item. Items 31–82 (age range approximately 1;0 to 3;6 year) were administered in 20% of the cases or more, which means more than 70 actual observations in the DD group. Items 1–18 and 90 and 91 had a very small number of observations in the DD group ($< 20$), which means that results for these items are for a large part based on automatically imputed scores and should therefore be interpreted with caution.

### 3.2. DIF

The results of the Automated Item Selection Procedure in R showed that 88 out of the 91 items formed a single scale, supporting a sufficient degree of unidimensionality for these items. Exceptions were items 1 and 2, which could not be included in the analysis because they did not contain any variance, and item 90, which appeared unscalable.

The results of the CDR-analyses and logistic regression are shown in Table 3. We identified an item as containing DIF when either of the methods yielded a significant result (using an overall alpha of 0.05, with Bonferroni correction within the CDR) with a high effect size (ETS $\Delta > 1.5$; Zwick, 2012). Items 1, 2, 4, 6, 7, 10 and 90 did not show any variance between participants within one of the groups and were therefore dropped from the analyses. A total number of 15 items were identified as expressing DIF, see Table 3.

The items identified as containing DIF deal with looking at objects or at a mirror, playing with and picking up objects, making puzzle boards, matching/grouping/discriminating, and play skills. Among both the identified and non-identified items, some do and

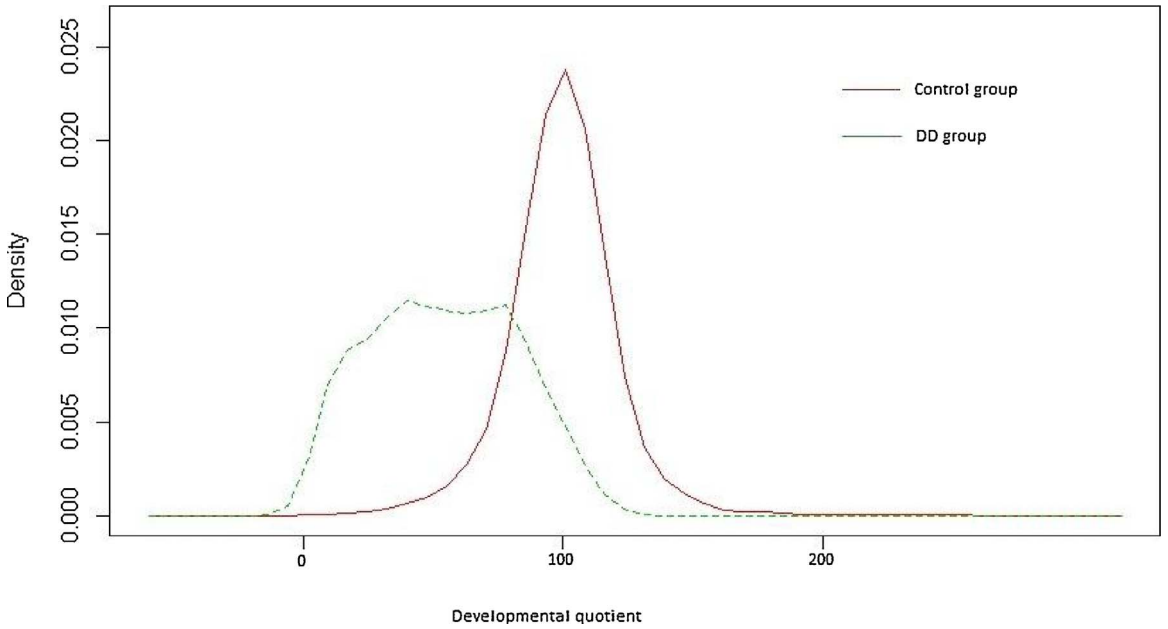


**Fig. 1.** Density plot of the Developmental quotient, per group.

**Table 3**
Results of the analyses for Differential item functioning.

| Item | Prop. Correct | CDR | | | | | | ETSΔ |
|---|---|---|---|---|---|---|---|---|
| | | MH | | | BD | | Log | |
| | | $\chi^2$ | | *ETSΔ* | $\chi^2$ | | *LRT* | |
| **3** | 1.00 | 0.16 | | −2.97 | 25.30 | ** | 7.14 | C |
| **22** | 0.85 | 1.93 | | −1.76 | 5.24 | * | 4.57 | C |
| **23** | 0.84 | 6.05 | * | −3.66 | 1.89 | | 6.49 | C |
| **27** | 0.80 | 10.96 | ** | −3.55 | 0.29 | | 14.05 | ** C |
| 30 | 0.77 | 0.47 | | −0.71 | 5.39 | * | 10.30 | ** A |
| **33** | 0.72 | 6.26 | * | −2.03 | 3.17 | | 21.99 | ** C |
| 34 | 0.70 | 0.06 | | −0.30 | 8.28 | ** | 12.36 | ** A |
| 35 | 0.69 | 0.81 | | −0.69 | 5.23 | * | 5.76 | A |
| 37 | 0.61 | 2.42 | | −0.95 | 11.37 | ** | 9.87 | ** A |
| 39 | 0.61 | 0.02 | | −0.01 | 10.59 | ** | 12.34 | ** A |
| 40 | 0.64 | 4.04 | | −1.48 | 0.14 | | 13.36 | ** B |
| 47 | 0.56 | 2.09 | | −1.36 | 105.79 | ** | 12.12 | ** B |
| **51** | 0.52 | 7.27 | * | 2.17 | 0.02 | | 8.38 | * C |
| 55 | 0.51 | 2.38 | | −1.14 | 25.94 | ** | 7.69 | * B |
| 57 | 0.41 | 5.93 | * | 1.40 | 0.13 | | 10.24 | ** B |
| **61** | 0.42 | 11.06 | ** | 2.66 | 2.32 | | 13.41 | ** C |
| **63** | 0.38 | 9.58 | ** | 2.01 | 0.03 | | 17.90 | ** C |
| **64** | 0.37 | 10.32 | ** | −1.84 | 19.25 | ** | 17.53 | ** C |
| 66 | 0.40 | 0.19 | | −0.38 | 7.69 | ** | 1.19 | A |
| **68** | 0.29 | 6.04 | * | 1.58 | 1.14 | | 8.05 | * C |
| **70** | 0.26 | 9.61 | ** | −1.65 | 45.63 | ** | 17.61 | ** C |
| **71** | 0.14 | 16.15 | ** | 1.97 | 1.41 | | 21.27 | ** C |
| **73** | 0.21 | 11.00 | ** | −2.14 | 2.12 | | 14.01 | ** C |
| 79 | 0.12 | 3.64 | | −1.26 | 42.65 | ** | 5.72 | B |
| **80** | 0.10 | 5.25 | * | −1.52 | 3.54 | | 5.14 | C |
| **91** | 0.00 | 1.01 | | 4.37 | 14.99 | ** | 2.35 | C |

*Note.* The table shows all items with at least one significant test result. DD = Developmental Disabilities, CDR = Combined Decision Rule, MH = Mantel-Haenzsel, BD = Breslow-Day Trend, * = $p < 0.025$, ** = $p < 0.01$, LRT = Likelihood Ratio Test, Classification of effect sizes: A = ETSΔ < 1.0, B = 1.0 < ETSΔ < 1.5, C = ETSΔ > 1.5. Bold item numbers mean that items were identified as containing DIF.

**Table 4**
Results of regression analysis predicting Guttman error ratio.

| Predictor | Estimate | Standard error | t-value | *p* |
|---|---|---|---|---|
| (Intercept) | 0.5741 | 0.0115 | 49.875 | < 0.000 ** |
| Developmental quotient | 0.0007 | 0.0006 | 1.322 | 0.186 |
| Group | 0.0854 | 0.0488 | 1.749 | 0.081 |
| Developmental quotient*Group | −0.0023 | 0.0010 | −2.224 | 0.026 * |

*Note.* Developmental quotient was centred at the population mean (=100).
* $p < 0.05$.
** $p < 0.01$.

some do not contain motor, visual, language, and social skills. Remarkably, for many of the series items, one item is identified and the others are not. Series items are items that have an identical administration procedure (which is only carried out once), but differ in their difficulty level.

### 3.3. Guttman errors

The average number of Guttman errors appeared 19.5 ($SD$ = 20.6, range 0–135) in the DD group and 14.6 ($SD$ = 13.9, range 0–116) in the control group. The difference between the groups is significant ($t$ = 4.27, $p < 0.001$). The average ratio of number of Guttman errors to the total number of administered items is also significantly higher in the DD group (0.73) than in the control group (0.57) ($t$ = 4.93, $p < 0.001$).

We performed multiple regression analyses to assess if the Guttman error ratio can be predicted by either of the variables group (DD or control), developmental quotient (centred), prematurity, motor, visual, auditory, or speech-language impairment, pervasive developmental disorder, Down syndrome, Phelan McDermid syndrome, and other genetic syndromes. As we suspected an interaction between group and developmental quotient after exploration of the data, we included this interaction in the analysis. The regression equation appeared significant ($F$(12,1832) = 2.68, $p < 0.01$, $R^2$ = 0.011), but none of the predictors appeared significant.

We performed a second multiple regression analysis with the predictors group and centred developmental quotient and their interaction. The regression equation appeared significant (($F$(3,1985) = 12.91, $p < 0.001$, $R^2$ = 0.018). The results are shown in

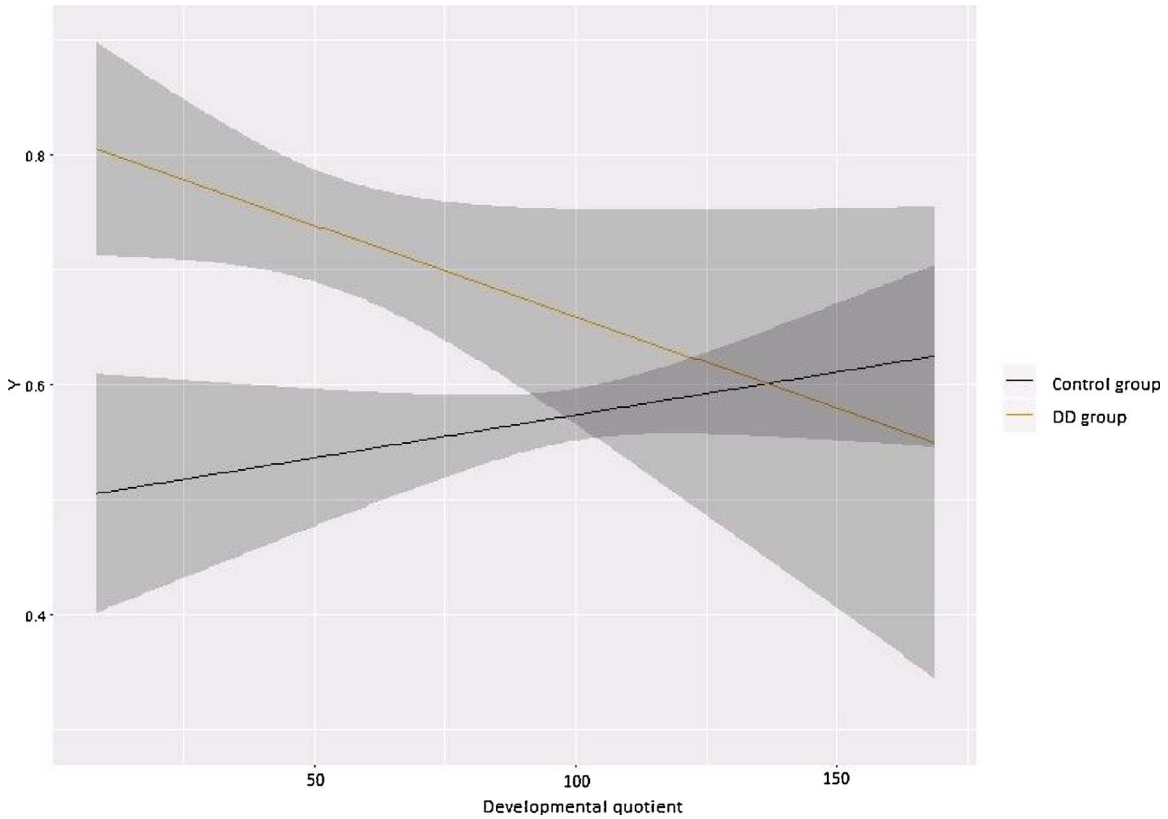


**Fig. 2.** Predicted Guttman error ratio (Y), as a function of Developmental quotient, per group.

Table 4. Only the interaction between developmental quotient and group appeared to predict the Guttman error ratio significantly. Fig. 2 was made on the basis of this second regression analysis. It shows the relationship between the Developmental quotient and the Guttman error ratio (Y) for the two groups. The DD group has a higher average Guttman error ratio. This is true for the whole range of Developmental quotient, given that the maximum of this range is 119 for the DD group. The lower the Developmental quotient, the higher the average Guttman error ratio in the DD group.

## 4. Discussion

The aim of this study was to examine whether qualitative differences exist in cognitive development of young children with and without DD, as measured by the Bayley-III Cognition items. We looked into DIF and the number of Guttman errors in children's test scores, as both form an indication of qualitative differences in development.

Our first hypothesis was that the development of children with DD deviates more from the typical order of skill development than the development of children without DD. The results support this hypothesis: 15 of the 91 items of the Bayley-III Cognition scale show DIF between the groups. The number of Guttman errors as a ratio to the total number of administered items per child appeared higher in children with DD. Combined, these results show qualitative differences in development between the groups and are in agreement with those of earlier studies into the use of the Bayley scale with children with DD, showing that results are not fully comparable to those with typically developing children (Milne, McDonald, & Comino, 2014; Moore et al., 2008; Wishart & Duffy, 1990). When taking into account the degree of developmental delay (operationalised as the developmental quotient), the presence of DD predicted the Guttman error ratio in an interaction with the developmental quotient, indicating that the differences in Guttman error between children with and without DD became larger with increasing degree of developmental delay. The results also show a number of Guttman errors ($M = 14.6$) within the control group that seems too high to be explained by measurement error and instability in performance alone. This means that a fixed order of skill development might not be an adequate assumption for children without DD either.

Our second hypothesis was that the groups would differ more with respect to scores on information processing tasks than on Piagetian tasks. This hypothesis is not supported: the items identified as having DIF did not measure a particular type of skill. Given that the presence of DIF did not even correspond within series items, it seems that variables other than the type of task are responsible for the DIF. These cannot be identified on the basis of the results of the current study.

Our third hypothesis was that children with DD due to organic causes and with a lower degree developmental delay would demonstrate most deviation from the typical order of skill development. The results do not support this hypothesis. The specific diagnosis or impairment of the child did not predict the Guttman error ratio. The Guttman error ratio did depend on the degree of developmental delay of the child. More specifically, contrary to our expectation, the lower the developmental quotient, the higher the average Guttman error ratio. It can be concluded that a deviation in the order of skill development is especially clear in children with DD with a low developmental quotient, which indicates that the likelihood of Guttman errors increases with the severity of DD.

A number of limitations need to be taken into account when interpreting the findings. First, development is a very complex process, based on a continuous interaction between the child and environment in interactive, transactional, and dynamic systems models of change (Sameroff, 2010). The test scores that form a basis for the current study depend on a large number of variables that influence the performance of a child in various ways. In the case of children younger than approximately two years, this complexity is increased due to the still undifferentiated character of development. The development in different domains is very strongly interrelated and therefore cognitive development can possibly not yet be regarded as an independent construct. In cases of DD, this complexity is increased further by the fact that this group is heterogeneous.

Second, due to following the prescribed procedure of the Bayley-III, we have administered only a subset of items per child and have imputed the other test scores. As items very low and high in the Cognition scale are administered relatively infrequently, the number of actually observed scores was low for these items. This means that our conclusions only count for the other items (19–89), and are more reliable for the items with a very high number of observations (ca. 30–80).

Third, the study into the Guttman error ratio is dependent on the order of the items in the Cognition scale. In the Dutch Bayley-III, and therefore the test administrations and analyses for the current study, the item order is the same as in the original American Bayley-III. The results of the Dutch standardisation study, however, show a deviating order, as is reflected in the indication of age levels per item in the Dutch scoring form. This means that we had possibly found different results if the item order had been changed accordingly in the Dutch version, which could partly explain why we found a relatively high Guttman error ratio for the control group as well. On the other hand, the different item order in the Dutch sample could also be caused by sampling fluctuations. The order might not be as fixed as the test suggests.

Coming back to the difference or delay paradigm, the results of the current study add to the research showing qualitative differences in cognitive development between children with and without DD. The results also show that not only children with DD, but also children without DD, show deviations from what is generally seen as a typical development. The question should therefore not only be if there is a difference between children with and without DD, but also if a typical order of skill development exists at all. Maybe this order is not universal, but just an average, with many deviations possible within the normal range.

More research is needed to identify how exactly the development of children with DD differs, for example in terms of specific skills and/or underlying processes. Preferably, future research into this topic should have a longitudinal design. Given the longitudinal character of development in itself, a longitudinal design is per definition better able to capture development (Weisz & Zigler, 1979) and differences between groups therein. Furthermore, more research is needed to estimate to what extend the assumption of the Bayley-III is problematic: the basal and ceiling rules allow for some deviation from the typical order of skill development. They

are determined such that no more than 5% of the children in the standardisation sample does not, respectively does, possess a skill below the basal or above the ceiling (Van Baar, Steenis, Verhoeven, Hessen, 2014). The results of the current study show that this percentage might be higher in cases of DD: the risk of overestimating the developmental level due to incorrectly counting skills below the basal, or underestimating due to missing skills above the ceiling, might be higher than 5%.

The results of the current study have consequences for the estimated validity of the Bayley-III. If we want to measure the skills of a child and compare these to the skills of children of the same age (e.g., standardised developmental assessment), and do this within a limited amount of time (Finello, 2011), we have to make assumptions. One of these assumptions is that there is a fixed order of development of skills within subtests. The conclusions of the current study cast doubt on this assumption. This means that a rigid adherence to the cognition score appears inappropriate. Because alternatives are often not available for estimating the developmental level of a child, the Bayley-III will remain a valuable instrument despite these limitations. However, the risk of missing skills should be kept in mind and dealt with, for example by additionally administering items below the basal and above the ceiling. This could best be implemented by defining broader basal and ceiling rules for cases in which the instrument is used to assess children with a high likelihood of large deviations in development (e.g., above 3;6 years of age or with severe DD). The quantitative results should then be regarded as a rough estimation of developmental level. The qualitative information on item performance should form the main focus. This information does not depend on the assumption of a fixed order in skill development and is most useful in formulating advice for the support of a child, as it allows the identification of strengths and needs. When depending on qualitative information, series items can be useful, because these assess skills at different levels.

As an estimate of the degree of developmental delay is generally used as a basis for diagnoses and decisions about support and services the child will receive, the Bayley-III Cognition score can have large implications for a child. This means that the risk of underestimating the developmental level should be taken seriously. The possibly qualitatively different development also needs to be taken into account in the support of a child. Adequate support needs to be given at the right moment (when a child is ready to develop a skill) and this moment could be at a different stage in the development than is the case in children with a typical development.

## 5. Conclusions

To conclude, Bayley-III Cognition items do show DIF when comparing children with and without DD, but the results of the current study do not explain this DIF. The order of skill development deviates from the typical order both for children with and without DD. This deviation is especially apparent for children with DD with a high degree of developmental delay.

## Acknowledgements

We would like to thank Prof. dr. B.F. van der Meulen, who sadly passed away in 2016. His mentoring in the preceding years and his ideas provided valuable support in writing this paper. Further, we would like to thank the research team at Utrecht University, who performed the standardisation research on the Bayley-III-NL, for sharing their research data.

## References

AAIDD (2008). *Frequently asked questions on intellectual disability and the AAIDD definition.* [Retrieved from http://aaidd.org/docs/default-source/sis-docs/aaiddfaqonid_template.pdf?sfvrsn=2].

AAIDD (2017). *Frequently asked questions on intellectual disability.* [Retrieved from https://aaidd.org/intellectual-disability/definition/faqs-on-intellectual-disability#.WLZ4g2_hCpo].

Acton, B. V., Biggs, W. S. G., Creighton, D. E., Penner, K. A. H., Switzer, H. N., Petrie Thomas, J. H., ... Robertson, C. M. T. (2011). Overestimating neurodevelopment using the Bayley-III after early complex cardiac surgery. *Pediatrics, 128*, e794–e800. http://dx.doi.org/10.1542/peds.2011-0331.

Bayley, N. (1969). *Bayley Scales of Infant Development.* New York: Psychological Corporation.

Bayley, N. (1993). *Bayley scales of infant development* (2nd ed.). San Antonio, TX: The Psychological Corporation.

Bayley, N. (2006). *Bayley scales of infant and toddler development* (3rd ed.). San Antonio, TX: Harcourt Assessment.

Bechger, T. M., Maris, G., & Verstralen, H. H. F. M. (2010). *Measurement and research department reports. A different view on DIF*Cito: Arnhem.

Bennett-Gates, D., & Zigler, E. (1998). Resolving the developmental-difference debate: An evaluation of the triarchic and systems theory models. In J. A. Burack, R. M. Hodapp, & E. Zigler (Eds.). *Handbook of mental retardation and development* (pp. 115–134). Cambridge: Cambridge University Press.

Bruckner, C., Yoder, P., Stone, W., & Saylor, M. (2007). Construct validity of the MCDI-I receptive vocabulary scale can be improved: Differential item functioning between toddlers with autism spectrum disorders and typically developing infants. *Journal of Speech, Language, and Hearing Research, 50*, 1631–1638. http://dx.doi.org/10.1044/1092-4388(2007/110).

Buntinx, W. H. E., & Schalock, R. L. (2010). Models of disability, quality of life, and individualized supports: Implications for professional practice in intellectual disability. *Journal of Policy and Practice in Intellectual Disabilities, 7*, 283–294.

Burack, J. A., Russo, N., Flores, H., Iarocci, G., & Zigler, E. (2012). The more you know the less you know, but that's OK: Developments in the developmental approach to intellectual disability. In J. A. Burack, R. M. Hodapp, G. Iarocci, & E. Zigler (Eds.). *The Oxford handbook of intellectual disability and development* (pp. 3–12). New York: Oxford University Press.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Centraal Bureau voor de Statistiek (CBS) (2011). *Statline.* [Retrieved from http://statline.cbs.nl/].

Couturier, G. L. G., & Tak, J. A. (2002). Diagnostiek bij kinderen jonger dan 6 jaar. [Assessment in children younger than 6 years]. In Th. Kievit, J. A. Tak, & J. D. Bosch (Eds.). *Handboek psychodiagnostiek voor de hulpverlening aan kinderen. [Handbook psychological assessment in the support for children* (pp. 625–679). Utrecht: De Tijdstroom.

Dykens, E. M., & Hodapp, R. M. (2001). Research in mental retardation: Toward an etiologic approach. *Journal of Child Psychology and Psychiatry, 42*, 49–71. http://dx.doi.org/10.1017/S0021963001006540.

Facon, B. (2008). A cross-sectional test of the similar-trajectory hypothesis among adults with mental retardation. *Research in Developmental Disabilities, 29*, 29–44. http://dx.doi.org/10.1016/j.ridd.2006.10.003.

Finch, H., Barton, K., & Meyer, P. (2009). Differential item functioning analysis for accommodated versus nonaccommodated students. *Educational Assessment, 14*, 38–56. http://dx.doi.org/10.1080/10627190902816264.

Finello, K. M. (2011). Collaboration in the assessment and diagnosis of preschoolers: Challenges and opportunities. *Psychology in the Schools, 48*, 442–453. http://dx.doi.org/10.1002/pits.20566.

Freeman, L., & Miller, A. (2001). Norm-referenced, criterion-referenced, and dynamic assessment: what exactly is the point? *Educational Psychology in Practice, 17*, 3–16. http://dx.doi.org/10.1080/02667360120039942.
Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement, 46*, 314–329. http://dx.doi.org/10.1111/j.1745-3984.2009.00083. x.
Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
Hallioglu, O., Gurer, G., Bozlu, G., Karpuz, D., Makharoblidze, K., & Okuyaz, C. (2015). Evaluation of neurodevelopment using Bayley-III in children with cyanotic or hemodynamically impaired congenital heart disease. *Congenital Heart Disease, 10*, 537–541.
Hasan, P. J., & Messer, D. J. (1997). Stability or instability in early cognitive abilities in children with Down's syndrome? *The British Journal of Developmental Disabilities, 43*, 93–107.
Hodapp, R. M., & Burack, J. A. (1990). What mental retardation teaches us about typical development: The examples of sequences, rates, and cross-domain relations. *Development and Psychopathology, 2*, 213–225.
Holland, P. W., & Thayer, D. T. (1985). *An alternate definition of the ETS delta scale of item difficulty (Research Report RR-85-43)*Princeton. NJ: Educational Testing Service.
Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel?Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.). *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.
Houwen, S., Visser, L., Van der Putten, A. A. J., & Vlaskamp, C. (2016). The interrelationships between motor, cognitive, and language development in children with and without intellectual and developmental disabilities. *Research in Developmental Disabilities, 53-54*, 19–31. http://dx.doi.org/10.1016/j.ridd.2016.01.012.
Komur, M., Ozen, S., Okuyaz, C., Makharoblıdze, K., & Erdogan, S. (2013). Neurodevelopment evaluation in children with congenital hypothyroidism by Bayley-III. *Brain and Development, 53*, 392–397. http://dx.doi.org/10.1016/j.braindev.2012.07.003.
Lautamo, T., Laakso, M.-L., Aro, T., Ahonen, T., & Törmäkangas, K. (2011). Validity of the play assessment for group settings: An evaluation of differential item functioning between children with specific language impairment and typically developing peers. *Australian Occupational Therapy Journal, 58*, 222–230. http://dx.doi.org/10.1111/j.1440-1630.2011.00941.x.
Lauteslager, P. E. M. (2000). *Children with Down's syndrome. Motor development and intervention.* Utrecht, the Netherlands: Utrecht University & Heeren Loo Zorggroep.
Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847–862. http://dx.doi.org/10.3758/BRM.42.3.847.
Magis, D., Beland, S., & Raiche, G. (2015). *Package difR.* [Retrieved from https://cran.r-project.org/web/packages/difR/difR.pdf].
Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*, 311–314.
Milne, S. L., McDonald, J. L., & Comino, E. J. (2014). Alternate scoring of the Bayley-III improves prediction of performance on Griffiths Mental Development Scales before school entry in preschoolers with developmental concerns. *Child: Care, Health and Development, 41*, 203–212. http://dx.doi.org/10.1111/cch.12177 Advance online publication.
Moore, D. G., Goodwin, J. E., & Oates, J. M. (2008). A modified version of the Bayley Scales of Infant Development-II for cognitive matching of infants with and without Down syndrome. *Journal of Intellectual Disability Research, 52*, 554–561. http://dx.doi.org/10.1111/j.1365-2788.2008.01064. x.
Morss, J. R. (1985). Early cognitive development: Difference or delay? In D. Lane, & B. Stratford (Eds.). *Current approaches to Down's syndrome*. London: Holt Rinehart & Winston.
Nabuzoka, D. (2008). Issues and developments in special education. In E. L. Grigorenko (Ed.). *Educating individuals with disabilities: IDEIA 2004 and beyond.* New York: Springer.
Pérez-Pereira, M., & Conti-Ramsden, G. (1999). *Language development and social interaction in blind children.* Hove, England: Psychology Press/Taylor & Francis (UK [Retrieved from https://books.google.nl/books?hl=en&lr=&id=9wd1xXuYwmwC&oi=fnd&pg=PP1&dq=development+blind+children&ots=MPKRIGngDj&sig=tlx75F4k4iJ9E8MdbWWASMOcCD4%20-%20v=onepage&q&f=false#v=onepage&q=development%20blind%20children&f=false].
R Core Team (2016). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing [Retrieved from https://www.R-project.org/].
Randall, J., Cheong, Y. F., & Engelhard, G., Jr. (2011). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement, 71*, 129–147. http://dx.doi.org/10.1177/0013164410391577.
Reimer, A. M., Cox, R. F., Boonstra, F. N., & Nijhuis-Van der Sanden, M. W. (2015). Measurement of fine-motor skills in young children with visual impairment. *Journal of Developmental and Physical Disabilities, 27*, 569–590. http://dx.doi.org/10.1007/s10882-015-9433-5.
Reuner, G., Fields, A. C., Wittke, A., Löpprich, M., & Pietz, J. (2013). Comparison of the developmental tests Bayley-III and Bayley-II in 7-month-old infants born preterm. *European Journal of Ppediatrics, 172*, 393–400. http://dx.doi.org/10.1007/s00431-012-1902-6.
Sameroff, A. (2010). A unified theory of development: A dialectic integration of nature and nurture. *Child Development, 81*, 6–22.
Intellectual disability. In R. L. Schalock, S. A. Borthwick-Duffy, V. J. Bradley, W. H. Buntinx, D. L. Coulter, E. M. Craig, & M. H. Yeager (Eds.). *Definition, classification, and systems of support (11th)*. Washington: American Association on Intellectual and Developmental Disabilities.
Spencer-Smith, M. M., Spittle, A. J., Lee, K. J., Doyle, L. W., & Anderson, P. J. (2015). Bayley-III cognitive and language scales in preterm children. *Pediatrics, 135*, e1258–e1265.
Stone, E., Cook, L., Cahalan-Laitusis, C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English-language arts assessment for students who are blind or visually impaired. *Applied Measurement in Education, 23*, 132–152. http://dx.doi.org/10.1080/08957341003673773.
Van Baar, A. L., Steenis, L. J. P., & Verhoeven, M. (2014). *Bayley-III-NL, afnamehandleiding [Bayley-III-NL, administration manual].* Amsterdam: Pearson Assessment and Information B.V.
Van Baar, A. L., Steenis, L. J. P., Verhoeven, M., & Hessen, D. J. (2014). *Bayley-III-NL, technische handleiding [Bayley-III-NL, technical manual].* Amsterdam: Pearson Assessment and Information B.V.
Van Braeckel, K., Butcher, P. R., Geuze, R. H., Van Duijn, M. A. J., Bos, A. F., & Bouma, A. (2010). Difference rather than delay in development of elementary visuomotor processes in children born preterm without cerebral palsy: A quasi-longitudinal study. *Neuropsychology, 24*, 90–100. http://dx.doi.org/10.1037/a0016804.
Van der Ark, A. L. (2012). New developments in mokken scale analysis in r. *Journal of Statistical, 48*, 1–27.
Van der Meulen, B. F., Ruiter, S. A. J., Lutje Spelberg, H. C., & Smrkovsky, M. (2002). *Bayley scales of infant development* (2nd ed − Dutch version). Amsterdam: Pearson Assessment and Information B.V.
Velikos, K., Soubasi, V., Michalettou, I., Sarafidis, K., Nakas, C., Papadopoulou, V., ... Drossou, V. (2015). Bayley-III scales at 12 months of corrected age in preterm infants: Patterns of developmental performance and correlations to environmental and biological influences. *Research in Developmental Disabilities, 45*(-46), 110–119. http://dx.doi.org/10.1016/j.ridd.2015.07.014.
Visser, L. (2014). *The Bayley-III-NL Special Needs Addition. A suitable developmental assessment instrument for young children with special needs.* Groningen: Stichting Kinderstudies.
Visser, L., Ruiter, S. A. J., Van der Meulen, B. F., Ruijssenaars, A. A. J. M., & Timmerman, M. E. (2013). Validity and suitability of the Bayley-III Low Motor/Vision version: A comparative study among young children with and without motor and/or visual impairments. *Research in Developmental Disabilities, 34*, 3736–3745. http://dx.doi.org/10.1016/j.ridd.2013.07.027.
Visser, L., Ruiter, S. A. J., Van der Meulen, B. F., Ruijssenaars, A. J. J. M., & Timmerman, M. E. (2015). Low verbal assessment with the Bayley-III. *Research in Developmental Disabilities, 36*, 230–243. http://dx.doi.org/10.1016/j.ridd.2014.09.014.
Wechsler, D. (2009). *Weschler preschool and primary scale of intelligence − Third edition: Dutch version.* Amsterdam: Pearson Assessment and Information B.V.
Weiss, B., Weisz, J. R., & Bromfield, R. (1986). Performance of retarded and nonretarded persons on information-processing tasks: Further tests of the similar structure hypothesis. *Psychological Bulletin, 100*, 157–175. http://dx.doi.org/10.1037/0033-2909.100.2.157.
Weisz, J. R., & Zigler, E. (1979). Cognitive development in retarded and nonretarded persons: Piagetian tests of the similar sequence hypothesis. *Psychological Bulletin, 86*, 831–851. http://dx.doi.org/10.1037/0033-2909.86.4.831.
Wishart, J. G., & Duffy, L. (1990). Instability of performance on cognitive tests in infants and young children with Down's Syndrome. *British Journal of Educational Psychology, 60*, 10–22. http://dx.doi.org/10.1111/j.2044-8279.1990.tb00918.x.
Wishart, J. G. (1993). The development of learning difficulties in children with Down's syndrome. *Journal of Intellectual Disability Research, 37*, 389–403. http://dx.doi.org/10.1111/j.1365-2788.1993. tb00882. x.
Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series,* i-30. http://dx.doi.org/10.1002/j.2333-8504.2012. tb02290. x.