Chapter 3

BSID-II: Correction for Prematurity

Abstract
The first revision of the Bayley Scales of Infant Development, the BSID-II, has been employed since 1993 in the United States and has been adapted worldwide. Since 2002 a Dutch translation and standardization is available in the Netherlands and Flanders (BSID-II-NL). A significant difference from the original Bayley test (BSID) is the classification of the scales in item sets with the corresponding basal and ceiling rules, which resulted in a new procedure for administration. Research suggests that there is a lack of uniformity in the application of this procedure, which may jeopardise the reliability of the test results. This chapter discusses the influence of the BSID-II administration procedure on the assessment of preterm children. Two suggestions for improvement are given. First the adapted Dutch version is discussed and how it may solve the problems resulting from the use of item sets and basal and ceiling rules. Second, an alternative administration procedure is put forward, that allows the use of norm tables, but takes into account the deviating developmental pattern of preterm children.

1 This chapter has been submitted for publication, slightly adjusted and it was co-authored by Bieuwe F. van der Meulen, Han Nakken and Henk C. Lutje Spelberg.
3.1 Introduction

Some years after its introduction the Bayley Scales of Infant Development – Second Edition (BSID-II; Bayley, 1993) became an internationally debated topic focusing on the uniformity of the administration of the BSID-II (Ross & Lawson, 1997; Matula & Aylward, 1997; Gauthier, Bauer, Messinger, Closius, 1999; Black & Matula, 2000; Glenn, Cunningham & Dayus, 2001; Chandlee, Heathfield, Damokosh, Salganik, & Radcliff, 2002). The aim of this paper is threefold. First, by a detailed discussion on the administration of the BSID-II with respect to the BSID (Bayley, 1969), we wish to contribute towards uniform and standardized use of the BSID-II, in order to obtain valid and reliable data on the development of infants at risk. Second, attention is given to the issue of prematurity correction in determining the starting item and to converting the raw scores to index scores. Prematurity correction entails the period of prematurity in weeks relevant to the expected date of birth. Third, the adapted basal and ceiling rules are discussed for the Mental and Motor Scale in the Dutch version of the BSID-II. In addition, an alternative administration procedure is suggested for testing preterm children. This chapter focuses on the use of the BSID-II for clinical assessments and for research purposes in measuring a child’s developmental course and intervention evaluation.

3.2 Important Modifications

To obtain a clear picture of the issues at hand, the most important modifications to the BSID-II testing procedure are discussed relative to the original BSID, in particular, the classification of item sets and the revised procedure to demarcate the basal and ceiling levels. These modifications have a major impact on the course of test administration and, as is apparent from the research literature, on the final results of the test.

Classification in Item Sets and the Basal and Ceiling Rules

The consequence of the wide age range of the BSID-II is that a substantial number of items prove too difficult or too easy when administered individually. In the
previous edition frustration and/or motivational problems might occur as a result of prolonged testing to determine the lower and upper limit. The lower limit for the Mental Scale was reached at 10 consecutive positive scores and the upper limit at 10 consecutive negative scores. Likewise, four positive scores and four negative scores determined the lower and upper limit for the Motor Scale. For reasons of time limitation (and frustration limitation 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 to make the content range of the item set large enough to fix the basal and ceiling value within a single item set, and also establishing sufficiently differentiated scores. Because of the classification of items into groups according to age, 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. For the Mental Scale, at least five items must receive positive scores in order to reach the basal level and three items are to be negative to determine the ceiling level. Should a child achieve fewer than five positive scores in the original item set, extra items are administered from a lower level and when a child performs well and has fewer than three negative scores, extra items are administered from a higher level. This procedure is continued until the basal level has been reached as well as the ceiling level (not necessarily in the same item set of the basal value). Only an exceptionally high or low score gives reason for administering items that are not included in the original item set. A consequence of this procedure is that a child will normally reach both ceiling and basal level within a single item set because of the broad range of the item set. All items prior to the first item of the item set are positive scores and all items following the last item are negative.

Consequences of Testing Preterm Children

Correction in determining the starting item. Research indicates that a child’s test result is partly explained by the item set the examiner starts with. The standard procedure is that both the item set at the start and the norms should correspond to the child’s
chronological age. However, in the case of a child with suspected developmental delay, the examiner may choose to start in the item set corresponding to the child’s level of functioning (Bayley, 1993; Van der Meulen, Ruiter, Lutje Spelberg, & Smrkovsky, 2002). If this procedure is followed for children whose development is considerably delayed, Washington, Scott, Johnson, Wendel, and Hay (1998) found that a child may receive several different Mental Developmental Indices, dependent on the item set at the start of the test. Washington and colleagues (1998) discussed cases of children who reached the basal and ceiling level in different item sets (between the item set at the start and the item set according to chronological age). Consequently, the results show different raw scores for the same child, dependent on the item set at the start of the test. When comparing the raw scores with the age norm, the outcome will show different Mental Developmental Indices. In some cases the differences in MDI are “alarmingly large”. Gauthier and colleagues (1999) supported this finding by concluding that the lower the selected item set at the start, the lower the index score will be. A group of at-risk infants of 12 months were tested with item sets for 11-, 12-, and 13-month-old children. Nearly all children reached the basal and ceiling level in all three item sets. When converting the raw scores it became clear that the lower the item set was at the start, the lower the Developmental Index (DI) when applying the 12-month-norm table. Twice as many infants would have received MDI’s of less than 85 using the 11-month set compared to the 12-month-set, moreover, MDI’s less than 85 were almost eliminated when the 13-month-set was used. Earlier results from research by Ross and Lawson (1997) reported that children with major developmental delays are likely to achieve similar scores whether one uses the chronological age or the adjusted age item set, but children who receive average scores when the chronological age item set is used are likely to have lower index scores if the corrected age item set is used. They add that these differences in scores could be substantial in individual cases and are most likely to occur in at-risk children with an uneven developmental course, the consequence of which is a higher risk of over or underestimation of the item set at the start of the test. As Glenn and colleagues (2001) stated, the problem is a result of the criteria for arriving at basal and ceiling levels. The children in above-mentioned studies could pass the basal criterion in higher sets yet fail enough items in a lower item set to reach the ceiling. In reaction to the study by Ross
and Lawson (1997), the developers of BSID-II emphasised in a report published later (Matula & Aylward, 1997) that the examiner should start in the item set matching the child’s chronological age or as close as possible when the child is developmentally delayed. They confirm that “the examiner should be conservative and try to select the item set closest to chronological age, because it is true that the item set will affect the score that the infant receives” (Matula & Aylward, 1997, p. 112).

**Correction in Converting Raw Scores to Standard Scores.** Justification can be found for the above-mentioned reaction by Matula and Aylward (1997), though it should be noted that the manual, as well as publications that followed, dictate that the age of the item set at the start should correspond with the norm table. This, for instance, applies to the use of the BSID-II with preterm children. If the examiner wishes to compare a child’s development with that of the average development of its age group the first step would be to proceed with the chronological age item set.

The first main target is to compare each child with the average development of its age group (normative reference group). The administration of the item set that corresponds with the norm table as to age, yields a comparison that is as fair as possible because both the child and normative reference group are presented with the same item set. The practice of testing shows that a large group of children are developmentally delayed to such an extent that age appropriate testing would arouse a great deal of frustration because of the high degree of difficulty of the items. This could produce a negative effect on the outcome of the test. The opposite may also occur; children who develop faster may experience motivational problems, as the items prove too easy. Both the American and the Dutch manual indicate that in some cases it is considered desirable to start in the item set below or above the chronological age closest to the child’s level of functioning. Special attention is given to the group of preterm children. Despite differing views on the development of preterm children, most will agree that these children show a delayed development compared to full term children. In most cases the delayed development will be taken into account. The starting item set plays an important role in determining developmental indices on the basis of both chronological and corrected age because of the structure of the BSID-II and its classification of scales into item sets. The items administered belong to a certain age group that should correspond with the age.
group of the norm table (Bayley, 1993; Van der Meulen et al., 2002) to reach a valid test result. The child’s chronological age serves as starting point for both the item set and norm table, but for a preterm child aged 19 months with a corrected age of 17 months, items that belong to the ‘interval’ age group between 17 and 19 months will be administered as well. Consequently, on the basis of such test administration, chronological as well as corrected age can serve as a starting point for converting the raw scores to the developmental index. A premature child born 12 weeks early who has reached a chronological age of 21 months on the day of testing and having a corrected age of 18 months for test administration will receive a group of items partially different from the ones that would have been administered in the item set of 20-22 months. The item set of 20-22 months on the Mental Scale includes eight extra items and the items below the basal level are positive scores. In this case it concerns 10 items. A child who is tested in the age range of 20-22 months set will receive credit for ten items for free and is therefore given an opportunity of scoring eight more items positively. As described earlier the clinician’s choice of the starting age-group will increase the chance of under and overestimation of the child’s developmental level (Washington et al., 1998).

3.3 Adaptations of the BSID-II-Dutch Version

The Bayley Scales of Infant Development-Second Edition-Dutch Version (BSID-II-NL, Van der Meulen et al., 2002; Ruiter, Lutje Spelberg, & Van der Meulen, 2005) is a translation and restandardization of the BSID-II for the Mental, Motor and the Behavior Rating Scale. In addition, a Non-Verbal Mental Scale was developed with separate standardization (derived from the original normative data) for the age range of 12 to 30 months (Ruiter, Hoekstra, Van der Meulen, & Lutje Spelberg, 2005).

General test administration procedure

In the essentials edition on the BSID-II (2000) Black and Matula (2000) reported that if an examiner wishes to compare the test results with the normative reference group, the test administration should follow a prescribed standardized procedure, without any special adaptations. Should the examiner, in addition to determining the general
developmental index also wish to collect data on for instance specific skills of the child or the process of information processing, he or she can adjust the administration procedure, although the norm tables for standard administration will no longer apply. The examiner will, on the other hand, still be able to include the extra information in the assessment and report on the child’s development.

An alternative test administration procedure for testing preterm children would allow for the use of norm tables and at the same time take into account the developmental pattern of these children. Such an alternative procedure enables the examiner to make a true comparison between a preterm child and the average development of members of its age group as well as its corrected age group. By administering the items belonging to both the chronological age group and the corrected age it is possible to determine two separate valid standard scores. One score realistically reflects the child’s development in comparison with the child’s age group members and the other score takes into account the preterm condition of the child. The item set at the start of this adapted test administration procedure with its altered ceiling and basal rules (see later in this chapter) is also expected to affect the test results. Our advice is to start with the items for the higher item set (belonging to chronological age), because this has been the procedure in the standardization study. There is less chance that the child becomes frustrated by the difficulty of the items because of the large overlap of the contents of the item groups and only a few extra items are required. Furthermore, numerous items are interrelated, the consequence of which is that the administration of one item will also yield the score for the next items in the series (or can be skipped in case of a negative score, such as, the ‘walking items’ on the Motor Scale). The examiner can incorporate successful experiences by alternating a number of difficult items with a number of easy, more appealing ones. This may require the use of additional item sets.

The basal and ceiling rules for both groups apply only when the raw scores are to be calculated. These scores are calculated for two separate tests, one at the chronological age level and the other at the corrected age level. This more flexible test procedure requires a thorough preparation of test administration. The examiner should determine the adapted order of the test items prior to the test. The recommended order of administration of the items from the first (and highest) item set can serve as a starting point. Information
from file analysis and from interviews with those involved help determine how items from the lower age groups should be integrated in the test administration. If the examiner is well prepared there need not be any negative consequences for the pace of test administration, contact with the child and final test result. If the child shows increasing resistance to the test as a result of fatigue or lack of motivation, which may jeopardize the validity of the test, the examiner may decide to postpone further administration. In that case, the second test administration is to be solely a continuation of the first test administration, excluding all items administered in the first test, unless the negative scores in the first test were clearly caused by fatigue or lack of motivation.

Preterm children often demonstrate not only a developmental delay but also atypical development; on average these children have lower scores on items concerning fine motor skills and concentration such as solving problems and items with multiple instructions. Black and Matula (2000) described how to estimate the regularity of a child's developmental course. Score development of an item set is studied, taking into account the child’s family history, previous test results and medical history. Most children will achieve positive scores on the easiest (first) items, show alternating positive and negative scores on the middle section and overall negative scores on the last and most difficult part of the item set. The scoring pattern will be less clear when children show an atypical development. In such case the examiner may decide to administer an adapted item set to test, e.g., the limits. By administering items from other age groups closest to the child’s age it is possible to test a certain skill at different levels (use can be made of appendix D of the BSID-II manual).

Adjusted basal and ceiling rules

An item set different from the one at the start is only used when the child’s achievements are far below or above the average age level. To retain the use of item sets and at the same time reduce the problem of ‘getting stuck’ in an item set, the Dutch version of the BSID-II has been adapted. 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 are to be
positive scores. 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 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 rules the limits of an item set will become less rigid; in this adapted version the child is less inclined 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 five or more negative scores. In doing so, 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.

It emerges from research data of a group of 18-month old preterm children, as described by Ruiter, Moens, Van der Meulen, Weisglas, Nakken, and Spelberg (2007), that when the American cut-off rule for the Mental Scale is applied, only 9.7% of the children failed to reach the item set of 17-19 months. After applying the Dutch basal level of at least eight positive scores, it appeared, however, that 23% of the subjects did not reach the basal and therefore had to go to a lower item set. The difference is less striking on the Motor Scale, 6.2 and 9.7 of the children respectively, lack sufficient positive scores to stay in the age group. There were no differences in the ceiling level, since all children reached the ceiling level in the item set of 17-19 months. Four cases are discussed in an article by Washington and colleagues (1998), three of which provide insight into the influence of the chosen item set and the application of the original basal rule for the Mental Scale. These cases concern children with a developmental delay or atypical development. In these cases the examiner may decide to start in an item set that does not correspond with the child’s chronological age but rather with the expected developmental level. It emerges from the table below that the developmental index shows variation according to the selected item set. This is the result of the different raw scores attained on the item sets. In case four, for instance, this difference amounts to 25 raw scoring points, as a consequence the developmental index diverges from <50 (at the start
in the item set 20-22 months) to 78 (at the start in the item set 32-34 months). If, however, the adjusted Dutch basal and ceiling rules are applied the differences will become less noticeable (difference is eight index points). As a result of applying the extended basal rule (> 8 positives) for the Mental Scale the differences in developmental index points for case two is 8 points (US 18; 61-79), case three: 0 points (US ≥8; 50-58) and case four: 8 points (US ≥28; 50-78).

Table 1

Test results comparing original (US) vs. adjusted (NL) basal and ceiling rules

<table>
<thead>
<tr>
<th>Item Set</th>
<th>Credit</th>
<th>No Credit</th>
<th>Raw Score</th>
<th>Mental Index US</th>
<th>Mental Index NL¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 2: Lauren</td>
<td>23-25 mo</td>
<td>14</td>
<td>22</td>
<td>126</td>
<td>61</td>
</tr>
<tr>
<td>29 mo</td>
<td>26-28 mo</td>
<td>8</td>
<td>24</td>
<td>130</td>
<td>69</td>
</tr>
<tr>
<td>29-31 mo</td>
<td>5</td>
<td>23</td>
<td>135</td>
<td>79</td>
<td>_-*</td>
</tr>
<tr>
<td>Case 3: Daren</td>
<td>26-28 mo</td>
<td>16</td>
<td>16</td>
<td>138</td>
<td>&lt;50</td>
</tr>
<tr>
<td>40 mo</td>
<td>29-31 mo</td>
<td>10</td>
<td>18</td>
<td>140</td>
<td>50</td>
</tr>
<tr>
<td>32-34 mo</td>
<td>7</td>
<td>24</td>
<td>142</td>
<td>54</td>
<td>_-*</td>
</tr>
<tr>
<td>35-37 mo</td>
<td>5</td>
<td>24</td>
<td>144</td>
<td>58</td>
<td>_-*</td>
</tr>
<tr>
<td>38-42 mo</td>
<td>4</td>
<td>29</td>
<td>146</td>
<td>_-*</td>
<td>_-*</td>
</tr>
<tr>
<td>Case 4: Ian</td>
<td>20-22 mo</td>
<td>9</td>
<td>20</td>
<td>115</td>
<td>&lt;50</td>
</tr>
<tr>
<td>33 mo</td>
<td>23-25 mo</td>
<td>11</td>
<td>25</td>
<td>123</td>
<td>51</td>
</tr>
<tr>
<td>26-28 mo</td>
<td>8</td>
<td>24</td>
<td>130</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>29-31 mo</td>
<td>5</td>
<td>23</td>
<td>135</td>
<td>68</td>
<td>_-*</td>
</tr>
<tr>
<td>32-34 mo</td>
<td>5</td>
<td>26</td>
<td>140</td>
<td>78</td>
<td>_-*</td>
</tr>
</tbody>
</table>

From: Washington et al. (1998): Table 1: Multiple Possible Outcomes for Cases 2, 3, and 4 (Mental Scale), slightly adjusted

¹ Mental index when applying the Dutch basal and ceiling level

* Mental Index can not be determined because basal level was not reached.
Developmental indices are important components in the decision making process concerning the possibilities and suitability of intervention programs and/or type of education. The government has created indication committees for the different types of education (clusters) and care. As the outcome of a test often plays a decisive role in the assessment of the (im)possibilities concerning the type of education or care, thus the child’s future, it is essential to employ the instruments in a correct and standardized way. This means that a standard instrument should be administered according to a standard procedure in terms of validity and reliability.

In principle, it should not make any difference which starting item set is administered of the BSID-II. The examiner should always arrive at the same raw score whether he or she starts in the chronological age range or corrected range. However, it does not always turn out that way. Especially children with a deviating developmental pattern will more often reach the basal level in consecutive age groups and the developmental index is more likely to be influenced by the selected item set. The extended basal and ceiling level offer a broader scope for administering the items from different age groups and consequently there will be less influence of the chosen item set on the final test result. The alternative test procedure enables the examiner to determine two raw scores that reflect the child’s developmental level in comparison to that of his chronological age group and also reflect the child’s developmental level, taking into account a potential or expected developmental delay of a preterm child.

This chapter discusses the correction for prematurity at two levels, for determining the starting item and for establishing the norm score. The discussion on the procedural aspects of the BSID-II for testing preterm children often focuses on guidelines to find out to what age a correction for prematurity is considered useful. We will briefly go into this question here. There is general agreement that even a premature infant with a benign neonatal course remains at increased risk of neurological morbidity (National Research Council and Institute of Medicine, 2000). In spite of current high technology that facilitates an excellent imitation of the mother womb, experts agree that a healthy child’s natural environment cannot be emulated. High technology all around, many different caretakers and being mostly separated from the parents makes this stressful
situation a traumatic experience for parent and child and can thus influence the child’s development (Weisglas-Kuperus, 1992). There are differing opinions on the application of correction for prematurity in assessing a child’s developmental course. Most studies show correction for prematurity for at least the first two years of the child’s life. In practice, this is also the convention. There are, however, researchers who advocate a correction for only the first year of a child’s life (Den Ouden, Rijken, Brand, Verloove-Vanhorick, & Ruys, 1991; Batshaw & Shapiro, 2002). Caputo, Goldstein, and Taub (1981) hold the opinion that predictions for future development are too optimistic if a correction for age is applied. It is their considered opinion that there is no need for correction. In a longitudinal study on the development of preterm children it would be more appropriate to start from non corrected scores, in order to find out whether there is real deficiency or temporary developmental delay. Studying the effect of interaction-intervention between preterm children and their mothers Schols (1993), using the BOS 2-30 (Van der Meulen & Smrkovsky, 1984) found that with correction the preterm children in the control group scored higher than the norm for full-term children at 15 and 24 months of age on the Motor Scale; their mental scores are equal to the norm. Without correction these children obtained mental scores that were below the norm at both 15 and 24 months; their motor development was equal to the norm at both times. Schols concluded from her study that a preterm child of 2 can safely be tested according to its chronological age; the extent of maturation hardly seems to make a difference at that stage. Other researchers support the notion of relating the correction for prematurity to the child’s birth weight (Hunt & Rhodes, 1977) or applying corrections for the various developmental domains (Siegel, 1983; Lems, Hopkins, & Samson, 1993; Palisano, 1986). Just as in the original version, the Dutch manual does not make a definite statement about the age up to which a correction for prematurity should be applied. The literature advises examiners to correct for maturity up to the child’s second year in case of severe prematurity. If prematurity is less severe there is more chance of overcorrection, considering the fact that these children often make up for their disadvantage in their first year of life.

Future research on the question of determining to which age correction for prematurity can be recommended should focus on examining the effect of prematurity on
test results. If test results become available for large numbers of children at different developmental levels, regression analysis can be applied to investigate the factor of prematurity on the test results at different ages. At the same time, as already stated by Ross and Lawson (1997), it would be important to further investigate the suggestion by Wilson (1987): continue reporting both scores until the corrected index score falls within the range of standard error deviation of the uncorrected index score.

The alternative test procedure for using the test with prematurely born infants has only been described in this chapter. The utility and empirical foundation of the alternative test administration procedure should however be demonstrated by providing longitudinal data that this change yields scores that are, in fact, more reliable and valid for this specific group of children.
References


Glenn, S. M., Cunningham, C. C., & Dayus, B. (2001). Comparison of the 1969 and 1993...


