

2. BAYLEY SHORT FORM–RESEARCH EDITION

As noted earlier, the design of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) was guided by three principles. The first guiding principle was to obtain measures of growth through repeated measures at multiple time points. The second was to obtain, wherever possible, direct measures of child functioning rather than to rely on parental reporting in order to reduce potential response bias. The third guiding principle was to obtain information about a broad spectrum of children's early experiences in order to understand their relationship to children's development over time.

As explained in section 1.6, there were to be two data collections during the toddler-to-preschooler period: one at 18 months and one at 30 months. Accordingly, an 18-month version of the Bayley Short Form–Research Edition (BSF-R) was developed and implemented in the 18-month field test, which began in May 2001. Simultaneous with this field test, design work was also conducted to identify a pool of candidate items for the 30-month version of the BSF-R. When the decision was made to combine these two data collections into a single data collection to occur when children turned 2 years of age, it was possible to benefit from work done on the 18- and 30-month versions of the BSF-R to develop the 2-year version of the BSF-R. Therefore, this chapter focuses on the work that was done leading up to the development of the 18-month version of the BSF-R mental scale and motor scale. Chapter 3 describes how the 18- and 30-month versions were used to form the basis of the 2-year BSF-R.

2.1 Decision to Use the Bayley Scales of Infant Development, Second Edition

In order to describe children's developing skills, it was necessary to select a measure of developmental status that provided a comprehensive snapshot of children's varying skills at multiple ages. In addition, because of the need for strong anchoring data points in the early childhood years, it was desirable to obtain a direct assessment of children's abilities rather than rely solely on parent reports. Parent reports can provide important converging evidence for children's abilities but do not substitute for direct assessments.

A screening instrument would be the most efficient measure to administer in the field setting of the ECLS-B. However, most screening instruments, such as the Bayley Infant Neurodevelopmental Screener or the Battelle Developmental Inventory are not comprehensive enough and do not offer the

breadth of developmental abilities desired for the ECLS-B; the items in such screeners represent behaviors and responses geared to the identification of pathology rather than the full range of developmental abilities. Since a key objective of the ECLS-B is to describe children's growth and development from infancy to the early school years, it was necessary to select a measure that provided a comprehensive snapshot of children's varying skills at multiple ages.

In addition to comprehensiveness, criteria for selecting an appropriate measure included the feasibility of field administration, the availability of well-standardized norms (to further anchor the study), reasonable predictive ability, the efficiency of administration, the age span of the measure, and its use in other large-scale studies.

The Bayley Scales of Infant Development, Second Edition (BSID-II) (Bayley 1993), described in more detail below, was found to fit the requirements of the ECLS-B on several levels. The BSID-II contains items appropriate from 1 month through 42 months of age. (The items are arranged in age sets so that only those items that are age-appropriate are administered.) As initially designed, it was the intention of the National Center for Education Statistics (NCES) to administer the full BSID-II, including the mental scale, motor scale, and the Behavior Rating Scale (BRS), to all sampled children in the ECLS-B at all data collections for which it was age appropriate.

Because the BSID-II could be administered at the 9-month and 2-year data collections, it would be possible to obtain continuity of measurement of growth in the ECLS-B. Previous studies have typically used a single BSID-II assessment to predict later child outcome results. The ECLS-B, by obtaining scores at two separate data collections, at 9 months and at 2 years, will enable analysts to examine developmental status as a repeated measure, which will help clarify the issue of the predictive validity of early childhood development assessments in general and the BSID-II in particular. The BSID-II was also selected because it offers the breadth of content that would provide the ECLS-B with a rich descriptive database to represent children's developing skills.

In addition, the BSID-II has the advantage of being among the more psychometrically rigorous standardized assessments available for infants and young children. It is generally recognized as the best assessment of developmental status available, in terms of reliability and validity, for children at this age. Critical psychometric properties of any standardized test include the precision of scores, stability of scores over time, and predictive validity. For further information about the psychometric properties and the rationale for selecting the BSID-II, please refer to the *ECLS-B Methodology Report for the Nine-*

Month Data Collection, Volume 1: Psychometric Characteristics (NCES 2005–100) (Andreassen and Fletcher 2005).

The BSID-II also has the advantage of having been used in other federally sponsored studies of early child development, such as the National Institute of Child Health and Development (NICHD) Early Child Care Study and the National Evaluation of Early Head Start. Using the BSID-II as the main baseline measure makes it possible to link the ECLS-B to those existing studies.

However, the excessive burden on interviewers and participants that was found in the fall 1999 9-month field test led to the decision to design shortened and streamlined versions of the BSID-II, for use at 9- and 18-months, the BSF-R.

2.1.1 Description of the BSID-II

The BSID-II is individually administered (i.e., one tester administers each item to one child) and assesses the current developmental functioning of infants and children from 1 month to 42 months. In total, the BSID-II is composed of two main scales, or sets of items: the mental scale and the motor scale. The mental scale consists of 178 items that assess abilities such as memory, habituation, problem solving, ability to vocalize, language, and social skills. The motor scale consists of 111 items that assess fine motor abilities, such as grasping and writing skills; and gross motor abilities, such as rolling, crawling and creeping, sitting, standing, walking, running, and jumping. All the items in the BSID-II are arranged in the order of their developmental difficulty. Most of the items must be administered, but a small percentage of them can be scored by observation during the administration of other items.

The BSID-II items are organized into age sets such that sets of items are administered depending on the child's chronological age. For example, the mental scale item set specified for a 24-month-old child includes 31 administered items, with 5 items scored by observation, for a total of 36 items. The motor scale item set specified for a 24-month-old includes 19 items, although 3 of those items could be combined into one administration with 3 scores. In the majority of cases, administration of the age-appropriate item set is sufficient to obtain an accurate assessment of a child's abilities. In some cases, however, it is necessary to administer additional sets of items to establish an accurate score. For children who do poorly and fail to score 5 or more *credits* within their item set on the mental scale, or 4 or more credits on the motor scale, the next younger item set is administered. For children who do very well and

score 3 or fewer *no credits* on the mental scale, or 2 or fewer *no credits* on the motor scale, within their age-appropriate item set, the next older item set is administered. Subsequent younger or older item sets continue to be administered until the basal or ceiling rule is satisfied.

According to the BSID-II manual, administration of the age-appropriate BSID-II at 2 years requires at least an hour to administer. Additional time is required if additional age item sets need to be administered to satisfy the basal or ceiling requirements.

Raw scores obtained from the number of passed and failed mental ability items and motor ability items are then converted, using look-up tables in the back of the manual, into a Mental Development Index (MDI) for the mental scale and a Psychomotor Development Index (PDI) for the motor scale. Both the MDI and the PDI have a mean of 100 and a standard deviation of 15, which places them on the same scale as many intelligence quotient (IQ) scores. Conceptually, however, the BSID-II should be thought of as an assessment of developmental status rather than of IQ. These index scores are normalized standard scores derived from a stratified quota sample based on U.S. Census figures for race/ethnicity, geographic region, and parent education. This standardization sample included only normal infants and children (children with physical problems, prematurity, medical complications, or developmental delay were not included in the standardization sample).

The BSID-II also includes a supplementary BRS, consisting of 30 items that assess the child's behavior during the assessment. The items comprise four facets according to age range: attention/arousal (1–5 months), and orientation/engagement, emotional regulation and motor quality (6–42 months). Examiners rate such aspects of the child's behavior as the child's interest in the test materials, soothability when upset, sociability, fearfulness, frustration with difficult tasks, and persistence. Scores on the BRS indicate the extent to which the child's behavior is considered within normal limits, questionable, or non-optimal for a child's age. Little information about the purpose and construction of the BRS is included in the BSID-II manual. Its most prevalent use is in clinical settings as an explanation for the child's performance on the mental and motor scales of the BSID-II. For example, poor performance on the mental scale could be due, at least in part, to frustration with difficult tasks or to poor emotional regulation.

2.1.2 Development of the BSF-R

Following the fall 1999 field test, members of the ECLS-B Technical Review Panel (TRP) were consulted about the production problems encountered during the field test. The following alternatives to the BSID-II were presented to the TRP: replace it with the Bayley Neurodevelopmental Screener; use a parent report measure such as the Minnesota Child Development Inventory (MN-CDI); drop the BSID-II at 9 months and at 2 years entirely; or administer either the BSID-II mental scale or the motor scale only at both time periods, or the motor scale at 9 months and the mental scale at 2 years. The consensus of the TRP was that a direct assessment of children's developmental status at 9 months and at 2 years was essential and that creation of an abbreviated version of the entire BSID-II for each of the data collection points was preferable to any of the other alternatives. They recommended using Item Response Theory (IRT) analyses to create an abbreviated version of the BSID-II because this technique makes it possible to add and subtract items without altering the underlying scale metric. (A brief overview of IRT analysis is presented in section 2.1.3.) This is the approach that was used, with considerable success, to develop the 9-month shortened BSID-II. For further information about the development of the 9-month BSF-R, please see the *ECLS-B Methodology Report for the Nine-Month Data Collection, Volume 1: Psychometric Characteristics* (NCES 2005–100) (Andreassen and Fletcher 2005).

In developing an abbreviated version of the BSID-II, it was necessary to ensure that it would maintain the psychometric properties of the original BSID-II and that it would successfully measure children's performance across the entire ability distribution, including the tails of the distribution. Selecting items on the basis of their face validity or the simplicity of materials would not be sufficient. IRT analysis would identify the items with the strongest psychometric properties for inclusion in the BSF-R. The assessment work group that guided the development of the 9-month BSF-R also guided the development of the 2-year BSF-R.

This work group consisted of four members, all of whom are experts in various aspects of assessment. Dr. Don Rock of the Educational Testing Service is an expert in IRT analysis and has extensive experience developing adaptive tests. He also served in this expert capacity on the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) and for the 9-month BSF-R for the ECLS-B. A second work group member, Dr. Kathleen Matula, is an expert in early child assessment who was involved in the restandardization of the BSID-II. The third work group member, Dr. Kathleen Williams of American Guidance Systems, is a psychometrician with extensive experience developing standardized assessments. The fourth member, Dr. Barbara Wasik, is an academic researcher in

developmental and educational psychology with extensive experience assessing cognitive development in low socioeconomic status (SES) and language minority samples. This work group reviewed Westat's IRT analyses for the development of the 18-month and 2-year forms of the BSF-R and provided comments on the results obtained from the 18-month field test and from the 9-month national data collection. Please see the *ECLS-B Methodology Report for the Nine-Month Data Collection, Volume 1: Psychometric Characteristics* (NCES 2005–100) (Andreassen and Fletcher 2005) for further information about the recommendations of this work group for the design of the BSF-R, specifically for the use of IRT 2-parameter logistic (2-PL) model and goals for reliability.

Additionally, members of the work group were consulted to ensure the quality of the data collection and administration of the measures. In addition to her participation in the assessment work group, Dr. Matula also served as an expert consultant on the administration and scoring of the items in the 18-month field test BSF-R and for the 2-year BSF-R. Prior to the 18-month field test, Dr. Matula conducted a 2-day training session for Westat's designated trainers for the 18-month training. After the redesign, she was consulted about any ambiguities in the administration steps and about the scoring of the 2-year items, for example, the number of trials permitted for "Builds tower of 6 blocks" and "Builds tower of 8 blocks," and whether they could be combined into a single administration. In addition, the BSF-R sections of both the 18-month and 2-year Child Activity Booklet were sent to Dr. Matula for her review to make sure that all items were accurately represented. To ensure consistency in the training of the approximately 200 interviewers for the national study, she also reviewed the accuracy of the 2-year BSF-R training videotape produced by Westat.

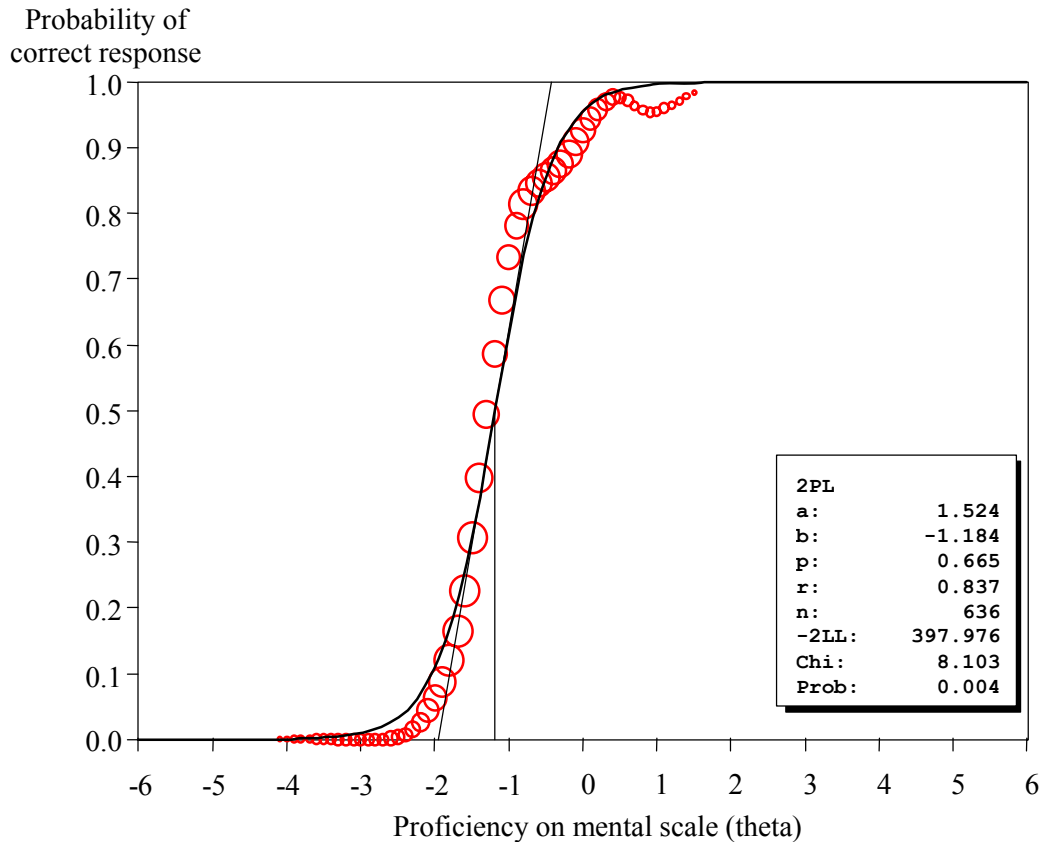
2.1.3 Overview of the 2-PL Response Model

IRT analysis is a powerful psychometric tool used in test construction and analysis.¹ The primary focus of IRT is the item response function, which models the probability of a correct response at different levels of ability. IRT analyses examine response data to generate item parameters used in scaling, scoring, and item selection. The ECLS-K battery was created using the 3-parameter logistic (3-PL) model, which includes an item difficulty parameter, an item discrimination parameter, and an item guessing parameter. Because the ECLS-B child assessment is not a multiple-choice test, there is no need for the 3-PL guessing parameter, and an IRT 2-PL model can be used instead. The 2-PL model includes only an item difficulty parameter and an item discrimination parameter.

¹ For additional information on Item Response Theory, see Baker (2001), which is available online at <http://edres.org/irt/>.

Exhibit 2-1 shows the response function or item characteristic curve (ICC) for a sample BSID-II item showing parameter values obtained with the publisher standardization dataset. The ICC, represented by a solid black line in exhibit 2-1, represents the probability $P_i(\theta)$ that a child with ability θ

Exhibit 2-1. Publisher item calibrations for a sample BSID-II item (MEN073, Turns pages of book), using publisher standardization dataset: 1993



NOTE: Mental item at level of difficulty $b = -1.184$, with power of discrimination $a = 1.524$. a =discrimination parameter; b =difficulty parameter; p =percentage correct; r =item-to-scale correlation; n =number of sample observations; $-2LL$ =-2 times Log Likelihood; Chi =Chi-square; $Prob$ =Significance of Chi-square; 2PL = 2-parametric logistic. Circles represent the empirical data and are proportional in size to the total number of observations at each point.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

will successfully complete item i . From a somewhat different perspective, this graph represents the proportion of children who will successfully complete this item at each level of ability. The response function is represented by an s-shaped curve that rises monotonically with ability between the limits of 0 and 1 over the ability range $[-\infty < \theta < \infty]$.

The formula for the 2-PL response function is:

$$P_i(\theta) \equiv \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} = \frac{1}{1+e^{-Da_i(\theta-b_i)}},$$

where

- Item difficulty parameter b_i acts as a location parameter, representing the point on the ability scale where the probability of a correct response is $p = 0.5$;
- Item discrimination parameter a_i acts as a slope parameter, determining the steepness of the response function's slope;
- Constant $D = 1.7$ is a scaling factor introduced so that the logistic function will resemble a normal ogive function as closely as possible, assuring that the 2-PL function will differ from the normal ogive function by less than 1 percent for all values of θ , and
- e is the exponential coefficient.

The probability of a correct response to an item in a given instance depends on the difference between the child's ability θ and item difficulty b_i . The greater the value of item difficulty b_i , the greater the ability θ usually required for a correct response. In relation to the scale at the bottom of exhibit 2-1, the ability distribution is centered around mean $\mu = 0$, with easy items located to the left of the mean, toward the low end of the ability distribution, and difficult items located to the right of the mean, at the high end of the ability distribution. For the particular item shown in this exhibit, the item difficulty parameter is $b_i = -1.184$. A more difficult item would be located to the right of this item (e.g., $b_i = 1.50$), and an easier item would be located to the left (e.g., $b_i = -2.50$). At 9 months, BSF-R mental items are generally found in the range from -3.0 to 1.0, and at 2 years in the range from 1.0 to 4.5.

The item discrimination parameter a_i is proportional to the slope of the ICC at b_i . Items with steeper slopes are generally more useful for making relevant distinctions of rank in children's ability levels near b_i . As the value of parameter a_i increases, the slope of the response function increases, increasing the amount of information provided by the item. As a_i decreases, the response function becomes flatter, and the item provides less information. Items with negative slopes are disallowed since this implies that the probability of a correct response decreases with ability. For this reason, parameter estimation is often based on the logarithm of a_i , which effectively avoids negative parameter values. Such items would normally be excluded from the scale.

Items with acceptable powers of discrimination will have item discrimination parameters in the general range of 0.7 to 1.0, with anything in the range of 1.0 to 2.0 generally considered to have especially high power of discrimination (Hambleton, Swaminathan, and Rogers 1991). Generally speaking, items with steeper slopes convey more information and yield ability estimates with smaller standard errors when tests are scored. An informative test will have an appropriate set of items with b_i and a_i item parameters, representing highly discriminating items distributed at strategic intervals across the ability distribution so that relevant distinctions of rank can be made. Unusually high discrimination parameter values, such as $a_i = 4.0$, are troublesome since this will usually show that the IRT assumption of local independence, conditional on ability, has been violated.

2.1.4 Creating the BSF-R for Round 2: Psychometric Rigor and Administrative Ease

Permission to create age-appropriate shortened versions of the BSID-II was sought and received from The Psychological Corporation, publisher of the BSID-II, which agreed also to call this shortened version the BSF-R. The Psychological Corporation also provided Westat with the standardization dataset for the BSID-II for the IRT analysis.

It should be kept in mind in the following discussion that the target age for the second round of data collection was initially set at 18 months and that it was subsequently shifted to 2 years. Therefore, discussion of the creation of the BSF-R for the second data collection begins with the 18-month version and follows the process through to the 2-year version, describing all work that was done.

Work toward developing the 9-month and 2-year BSF-R was guided by two considerations: psychometric rigor and administrative ease. Psychometric rigor was obtained through IRT analysis to ensure that the psychometrically strongest items were included. These analyses are described in depth in chapters 3 and 4.

The location of the target-age population on the ability distribution was identified. To oversimplify a bit, one function of IRT analysis is to line up all the items according to their ability level. Ideally, the items will line up at evenly spaced intervals across the entire ability range. Using the publisher's standardization dataset, the ability distribution appropriate for the 18-month field test (i.e., 17 to 19 months) was identified and then extended a bit at each end to take into account any children born prematurely and those children who might be assessed at a later age. It was also important to obtain good

measures for children located at the tails of the ability distribution. As a result, the ability distribution for the 18-month data collection BSF-R mental scale ranges from -0.458 to 6.76 population standard deviations (where the 12-month population² has a mean of 0 and a standard deviation of 1, which corresponds to an item age range from 9 to 37 months). The ability distribution for the 18-month BSF-R motor scale ranges from -0.773 to 5.367 standard deviations, with an item age range from 8- to 42-months. Working within this ability range, items were selected at approximately equal intervals along the ability distribution. Ideally, the criterion for selecting an item was an IRT discrimination parameter value of 1.0 or higher, although as low as 0.7 was considered acceptable. Values below 0.7 were avoided unless there were no higher values within that given range of difficulty. For example, given three items with ability parameters of 1.10, 1.21, and 1.24 and discrimination parameters of 0.6, 0.4, and 0.5, respectively, the item with the discrimination parameter of 0.6 would be selected in order to have an item that represented that range of ability. In addition, items were deleted on the basis of redundancy of coverage—if two items represented the same construct, say *means-end problem solving*, and had similar difficulty values, the one with the lower discrimination value was dropped if ease of administration was roughly equal.

The next step, after eliminating the psychometrically weak and redundant items, was to focus on administrative ease and include only those items that could reasonably be administered in a field setting by field interviewers. Items also must have had relatively objective scoring criteria. Administrative selection criteria were formulated to complement the IRT analytic criteria, as described below.

Minimal materials. Minimizing the number of materials needed was an important consideration. For example, the item “Identifies objects in photograph” requires a stimulus tray with preformed insets in which to place a rabbit, bell, block, car, and a small triangle, a shield to obstruct the child’s view of the tester arranging the materials, and the stimulus booklet, a spiral-bound book of about 50 pages that contains visual displays that are necessary supplements for some items. Similarly, several step-climbing items on the motor scale require that the interviewer tote a small set of steps built to specific standards. The ECLS-B interviewers have about 25 pounds of equipment to carry, including laptops, physical measurement equipment, and video cameras. Anything that could be done to reduce the number of BSF-R materials was desirable, and toting a small set of steps was not feasible. Therefore, these items were not included.

² Twelve months was selected as the reference point because it is in the center of the publisher’s sample in terms of number of observations.

Administration difficulty. Items that were difficult to administer were targeted for deletion. For example, the above-mentioned “Identifies objects in photograph” not only involves multiple materials but is also time-consuming and complicated to administer and, therefore, complicated to train interviewers to do. First the administrator places each object (bell, rabbit, block, triangle, car) on the tray according to the photograph in the stimulus book. The tray is then placed 9 inches in front of the child so that the car and cube are closest to the child. The administrator then points to the rabbit and says, “What is this?” If the child responds “rabbit” (or any appropriate name, such as “bunny”), then the administrator hides the tray from the child’s view with the shield and presents the photograph of the object tray from the stimulus book (in the same orientation as the actual object tray) and says to the child, “Show me the rabbit in this picture.” This process is repeated for the bell, cube, car, and triangle. The child receives credit for identifying at least two of the objects (although all five objects must be administered).

Objectivity of scoring. It was also desirable to exclude items with difficult or subjective scoring criteria. For example, the item “Makes a contingent utterance” requires that the administrator make a judgment about whether a child’s verbalization was in response to the speech of another individual, (e.g., the mother), or was produced independently of another’s speech. As a rule, the ECLS-B interviewers, most of whom were untrained in child development, early childhood education, or testing, had difficulty making inferences about children’s intentionality during verbal and behavioral responding. Therefore, item-scoring criteria needed to be as objective as possible so that interviewers would know what to observe. Items that were too subjective to score were excluded.

Maximize “twofers.” In the BSID-II, it is sometimes the case that multiple scores can be obtained from one administration. For example, “Builds a tower of 2 cubes,” “Builds a tower of 6 cubes,” and “Builds a tower of 8 cubes” have the same instructions and materials. The child is told to use all the cubes and “build a tower as big as you can.” Therefore, all three items can be scored from the same administration. A child who builds a tower of 4 cubes would receive credit for Builds a tower of 2 cubes and no credit for 6 cubes or 8 cubes. Within the constraints imposed by the psychometric power of the items, as many multiple scores from a single administration were included as possible. From an administrative viewpoint, this was an advantage. However, from the viewpoint of IRT, this was a disadvantage because it introduced the problem of interdependence of items. This was handled analytically during the IRT analyses, discussed in greater detail below.

Breadth of content. An additional goal was to maintain as much of the content of the items as possible. The BSID is atheoretical and is based on the author's observations of numerous children's abilities, incorporating successful items culled from other assessments, such as the Gesell Developmental Schedules (Gesell 1949). To the extent that it was possible, items were selected to capture as much of the content range as possible provided that an item had adequate psychometric properties.

2.1.5 IRT Analysis and an Adaptive Testing Strategy

Similar to the 9-month BSF-R, IRT principles were used to develop a BSF-R at 18 months that compared as closely as possible to publisher standards. One of the advantages of IRT is that items can be added to or deleted from a test while preserving the same scale metric. When response data are shown to satisfy IRT assumptions, item and ability parameters are sample free. Different samples of people yield the same item parameters. Different subsets of items yield the same ability parameters. The same results are obtained in every instance, implying that the measurement process is objective, external to either the specific set of items or the people encountered on any testing occasion.

Strictly speaking, tests with different numbers of items cannot be considered parallel forms, due to differences in test reliabilities. Although such tests fail to satisfy rigorous requirements for test equating, when data satisfy IRT principles, tests based on the same item pool can be calibrated on a common scale. These tests will then yield ability estimates for individuals that have the same central tendency but different standard errors. Tests drawn from the same item pool will then provide unbiased estimates of ability, although longer tests will usually provide more precise estimates. IRT offers the prospect of providing comparable scores that share the same scale metric found in publisher data.

The 2-year BSF-R was designed with IRT techniques to produce results that are as consistent as possible with those obtained using the BSID-II at this age range. The BSF-R diverges from the BSID-II primarily in its use of shortened core, basal, and ceiling item sets. The standard of comparison remains the BSID-II, based on the full complement of age item sets administered to children in a clinical setting. For the ECLS-B, the BSF-R is specially adapted for home administration as part of a household interview survey while replicating, as closely as possible, results that would be obtained using the full BSID-II. The use of items with high values on the discrimination parameters in each of the reduced item sets helps ensure measurement precision across the full range of the target population ability distribution.

There were four steps in developing the BSF-R, which are summarized as follows:

1. IRT calibration of the full complement of 178 mental and 111 motor items comprising the BSID-II mental and motor scales, respectively, using a 2-PL IRT model and the publisher standardization dataset.
2. Consulting publisher IRT item difficulty and discrimination parameters to select optimal subsets of core, basal, and ceiling items for the BSF-R.
3. Field testing BSF-R instruments, field test item calibrations, trial IRT true score equating with publisher tests, and reformulation of BSF-R instruments based on comparisons with BSID-II item calibrations.
4. Final BSF-R item calibrations, using the ECLS-B 2-year national dataset, final IRT true score equating using the publisher test as the target, generation of ability estimates, and indices of child development reported in publisher scale metrics.

Each of these steps is described in the following sections of this report. Readers familiar with IRT analysis and those who are familiar with the *ECLS-B Methodology Report for the Nine-Month Data Collection, Volume 1: Psychometric Characteristics* (NCES 2005-100) (Andreassen and Fletcher 2005) may wish to skim section 2.1.6 because the general content is redundant with similar material in the 9-month report.

2.1.6 IRT Item Calibrations of the BSID-II Standardization Dataset

The BSID-II includes 178 mental and 111 motor items designed for children between 1 and 42 months of age that are administered in age sets to avoid frustrating a child with items that are developmentally inappropriate. Basal and ceiling rules are devised to determine whether it is necessary to test outside the range of the designated age item set. Taking advantage of the large number of original BSID-II items, it is possible to shorten administration of the BSID-II by using smaller item subsets and an adaptive testing strategy.³

One objective of an adaptive testing strategy is to develop a core item set that is appropriate for most of the children in the target age group. The raw score total for these core items can then be used

³ In a traditional test, all individuals receive all the items in the test. In adaptive testing, however, the individual's performance on the first set of items determines whether, and which, additional item sets are administered. Individuals performing above some predetermined criterion (e.g., 1 standard deviation above the mean) would be routed to more difficult ceiling items whereas individuals performing below criterion (e.g., 1 standard deviation below the mean) would be routed to easier basal items.

to determine whether any specific child should be administered additional basal or ceiling item sets. The general idea is to test the limits of each child's ability with the recommended age item set, followed by the administration of additional basal and ceiling item sets as needed. When these additional items are required, all of the items in the supplementary item set are to be administered. Indeed, this adaptive strategy closely parallels the standard procedures of administration recommended by the publisher of the BSID-II. Since adjacent item sets contain overlapping items, this usually requires administering 4 to 10 items for each additional age item set.

IRT has been developed to represent item characteristics that result when an examinee encounters an item on a test. Item response models postulate that the probability of a correct response to an item on a test is a function of the difficulty of the item and the ability of the examinee. Assuming that all items represent the same ability domain, difficult items will be answered correctly less often than easy items. Given the difficulty of the item, more able examinees will provide a correct response more often than less able examinees.

The ICC represents the probability of a correct response in relation to examinee ability and item difficulty. Considering a single item, examinees at progressively higher levels of ability will have increasingly higher probabilities of a correct response. Alternatively, by considering a single examinee, items at progressively higher levels of difficulty will have increasingly lower probabilities of a correct response.

The probability of a successful outcome rises with examinee ability and falls as item difficulty increases. The outcome is governed by the difference between examinee ability and item difficulty in a specific instance. An incorrect response is more likely when examinee ability falls short of item difficulty; the odds of a correct response are even when examinee ability equals item difficulty; and a correct response is more likely when examinee ability exceeds item difficulty.

One feature of IRT is that examinee ability and item difficulty share the same scale metric. Examinees and items can be plotted opposite one another along the same scale axis. This implies that examinees can be represented by items at the appropriate level of difficulty, and items can be represented by specific kinds of examinees. Ability levels can be expressed in terms of the kinds of items that an examinee is able to complete successfully. Similarly, by observing examinee outcomes on a set of items, it is possible to work backward and infer the examinee's level of ability.

The ICC is a monotonically increasing function that represents the probability of a correct response at different levels of ability. The mathematical form of this function depends on the item—especially on how the item is scored. The BSID-II is based on a series of items representing child behavior. Instead of answering items on a test, as older children do at school, child behavior is observed on a series of specific tasks presented by an examiner. Item responses are based on the examiner's perception of the child's behavior as he or she attempts to undertake each task.

The examiner records whether or not the child is able to complete the task successfully. These observations are analogous to the credit-no credit scoring of questions on a test at school. In the case of the BSID-II, there are only two outcomes of interest. The child is presented a task to perform. The outcome is either successful or not, with little or no opportunity for guessing, much like a correct or incorrect response to a constructed-response item on a test.

Examiner observations of child behavior provide the basis for developing an item response model that represents the probability of successfully completing a task as a function of the difficulty of the task and the ability of the child. In IRT, a 2-PL response model is used to represent dichotomous outcomes of this type.

The 2-PL model features an item difficulty parameter b , which determines the location of the ICC on the ability axis, together with an item discrimination parameter a , which determines the rate of increase or slope of the ICC as ability rises. By examining the item parameters, it is easy to determine the relative difficulty of items and to determine which items are most discriminating⁴ at each ability level. Parameter estimation is referred to as *item calibration* and involves fitting the ICCs to the actual item responses. Parameter estimates are selected that maximize the likelihood of item responses across all ability levels for the sample as a whole. The likelihood of ability estimates θ is calculated concurrently as part of the item parameter estimation cycle. Several iterations of estimation and likelihood maximization are required before parameter values converge to yield a stable set of item calibrations.

The item response model is used to assess item format and the overall quality of the scale. After issues of scale reliability and validity have been addressed, scale scores and standard errors of measurement are generated to represent each infant's level of development. These scale scores enable the analyst to examine substantive issues of infant development.

⁴ That is, how successfully the item distinguishes between ability levels below which the individual received credit and ability levels above which the individual did not receive credit.

A sample of actual item responses is required for calibration purposes. Publisher data affords this opportunity. The BSID-II was developed by The Psychological Corporation by observing a combined sample of 2,939 children under clinical conditions. The combined sample includes a standardization sample of 1,700 observations of normal children, arranged in 17 age groups (ranging from 1 to 42 months of age, by month from 1 through 6 months, bimonthly from 6 through 12 months, trimonthly from 12 through 30 months, and semiannually from 30 to 42 months) and 1,239 additional observations. This information has been used by the publisher to develop an ordered listing of number-right raw scores for each age group, together with a corresponding set of standardized index scores that allow the comparison of developmental status among children of different ages. Standardized developmental index scores (T-scores in ECLS-B) are number-right raw scores that have been normed for each of several age groups. Publisher developmental index scores for BSID-II have a mean of 100 and standard deviation of 15 in each age group. T-scores in ECLS-B have a mean of 50 and standard deviation of 10 in each age group. The standardization sample contained 100 observations for each of the 17 selected age groups (table 2-1).

Table 2-1. BSID-II standardization sample: Mental and motor raw scores and index scores means and standard deviations, by age group: 1993

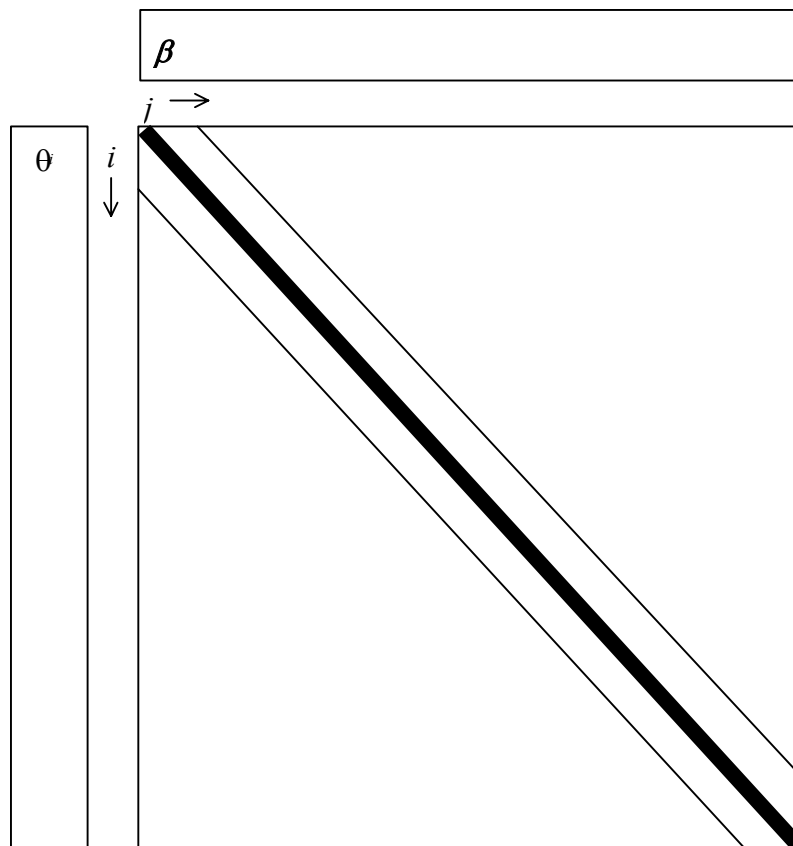
| Months of age | Sample N | Mental scale | | | | Motor scale | | | |
|---------------|----------|--------------|------|-------------|------|-------------|-----|-------------|------|
| | | Raw score | | Index score | | Raw score | | Index score | |
| | | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 1 | 100 | 15.3 | 9.4 | 101.8 | 18.2 | 11.7 | 3.9 | 101.5 | 13.5 |
| 2 | 100 | 27.4 | 7.2 | 99.8 | 14.7 | 16.5 | 5.4 | 100.0 | 15.0 |
| 3 | 100 | 33.5 | 7.9 | 100.0 | 15.7 | 25.0 | 7.0 | 99.4 | 19.6 |
| 4 | 100 | 44.5 | 7.7 | 99.9 | 15.4 | 28.6 | 6.3 | 99.5 | 18.2 |
| 5 | 100 | 55.4 | 7.7 | 99.9 | 15.0 | 33.5 | 4.3 | 99.5 | 14.5 |
| 6 | 100 | 62.8 | 7.0 | 100.3 | 14.9 | 39.9 | 5.7 | 100.3 | 17.4 |
| 8 | 100 | 71.9 | 6.8 | 100.8 | 14.8 | 53.3 | 5.3 | 99.7 | 15.5 |
| 10 | 100 | 78.3 | 4.7 | 99.5 | 10.6 | 58.1 | 3.5 | 101.4 | 12.9 |
| 12 | 100 | 87.7 | 6.6 | 100.2 | 15.3 | 64.6 | 3.9 | 99.5 | 15.7 |
| 15 | 100 | 98.4 | 5.9 | 99.7 | 11.8 | 69.5 | 4.0 | 99.0 | 16.2 |
| 18 | 100 | 112.4 | 9.0 | 99.6 | 17.2 | 75.3 | 3.4 | 100.2 | 13.4 |
| 21 | 100 | 123.8 | 8.8 | 99.6 | 17.2 | 78.6 | 3.6 | 98.8 | 13.8 |
| 24 | 100 | 132.9 | 9.6 | 99.5 | 18.1 | 83.9 | 4.1 | 98.8 | 15.3 |
| 27 | 100 | 141.4 | 10.1 | 99.8 | 19.8 | 90.4 | 5.7 | 100.7 | 19.2 |
| 30 | 100 | 146.6 | 6.8 | 99.5 | 14.2 | 93.6 | 3.5 | 100.5 | 13.4 |
| 36 | 100 | 155.4 | 7.4 | 100.8 | 14.9 | 100.1 | 4.0 | 100.3 | 14.5 |
| 42 | 100 | 165.1 | 7.3 | 100.2 | 14.5 | 105.2 | 3.1 | 101.3 | 13.0 |

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Specific age item sets are recommended for age groups between 1 and 42 months of age, with an average of 28 items in each set. Every age item set contains items that belong to more than one item set and thus overlaps with and provides linkages to adjacent age item sets. Sorting observations and items by age, valid item responses fall along a diagonal extending from the upper left to lower right of the data matrix. The thick diagonal line in figure 2-1 represents the core item sets recommended for adjacent age groups, with limited overlap in basal and ceiling items linking adjacent core item sets.

Parallel lines to either side of this diagonal line represent the additional basal and ceiling items that may apply in a given instance, depending on a child's level of development. The basal items for one age will generally belong to the item age set recommended for a previous age group. Likewise,

Figure 2-1. Schematic representation of publisher data, Item Response Theory ability estimates θ_i , and item parameters β_j : 1993



NOTE: i = rows of individuals sorted by individual ability (θ_i) and j = columns of items sorted by item difficulty (β_j).

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

ceiling items for one age will often include items from the age set recommended for subsequent ages. Thus, for a limited number of children with exceptional levels of development, basal and ceiling items provide additional overlap linking adjacent age item sets. Among observations in the standardization sample, 8.9 percent of the infants were administered basal items, while 14.1 percent received ceiling items.

The 1,700 observations in the standardization sample are complemented by an additional 1,239 observations of other infants having the same general demographic characteristics. Among these complementary observations, 13.5 percent were administered basal items, while 7.8 percent received ceiling items. The higher percentage of basal administrations suggests that perhaps 4.5 percent of the children in this second set of observations show evidence of disability. For scaling purposes, it is appropriate to take advantage of the larger number of observations in the combined sample of 2,939. This affords a larger number of item responses linking adjacent age item sets.

Common item linkages are used to calibrate the full set of BSID-II items on the mental and motor scales spanning development between 1 and 42 months of age. Item calibrations require that a latent population distribution be chosen to establish an IRT metric for ability and difficulty parameters. The origin and scale of the latent ability distribution is arbitrary. The convention is to calibrate items assuming a standard normal $N(0,1)$ distribution for latent ability, with population mean $\mu = 0$ and standard deviation $\sigma = 1$ (Hambleton, Swaminathan, and Rogers 1991).

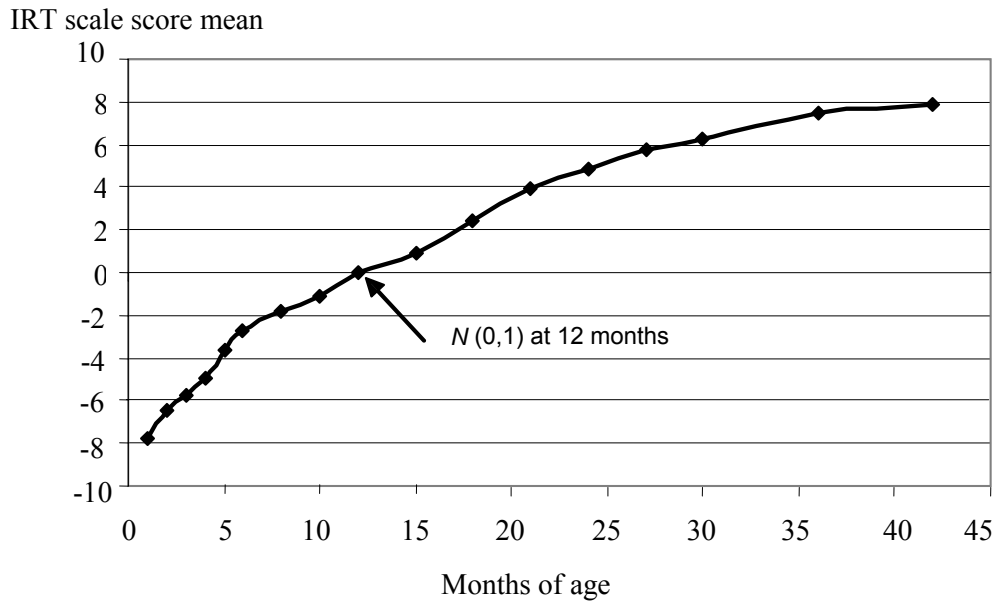
An age group at the center of the sample age distribution was selected to establish the origin and scale for the BSID-II IRT metric. The latent ability distribution of the 12-month age group was selected to have mean $\mu = 0$ and standard deviation $\sigma = 1$ on both the mental and motor development scales. This does not make the mental and motor scales directly comparable; it only establishes the 12-month age group as a common reference population.

Bilog-MG (Zimowski et al. 1996) and in-house software were used during item calibration and produced essentially identical parameter estimates. Both programs use marginal maximum likelihood estimation and allow the latent group population densities to be estimated concurrently with the item parameters. A multigroup IRT model was used, with observations clustered by age group. Common item linkages define the means and standard deviations for the 17 age groups in the sample by using the 12-month age group as a reference population. Working outward from the scale's origin at 12 months of age,

items and age group populations find their respective positions along a common development scale as part of the item calibration process.

Since mental and motor growth in early childhood is quite explosive (i.e., rapidly accelerating), the resulting development scales span many population standard deviations between 1 and 42 months of age. For the mental scale, estimated population means for the different age groups range between $-8 < \theta < 8$ population standard deviations, as shown in figure 2-2. The IRT scale is considered to be a true interval scale, implying that a unit increment at any point in the scale will represent an equivalent amount of relative effort. The IRT scale shows that early child growth is explosive and slows with advancing age. That is, between the mean at 1 month and the mean at 42 months, children will progress 16 population standard deviations. The first 8 standard deviations are passed by 12 months of age. The last 8 standard deviations take another 30 months of age. This shows that growth is especially rapid in the first year of life and then slows with age.

Figure 2-2. BSID-II mental scale score means by age: Item Response Theory (IRT) 2-parameter logistic item calibrations using publisher data: 1993

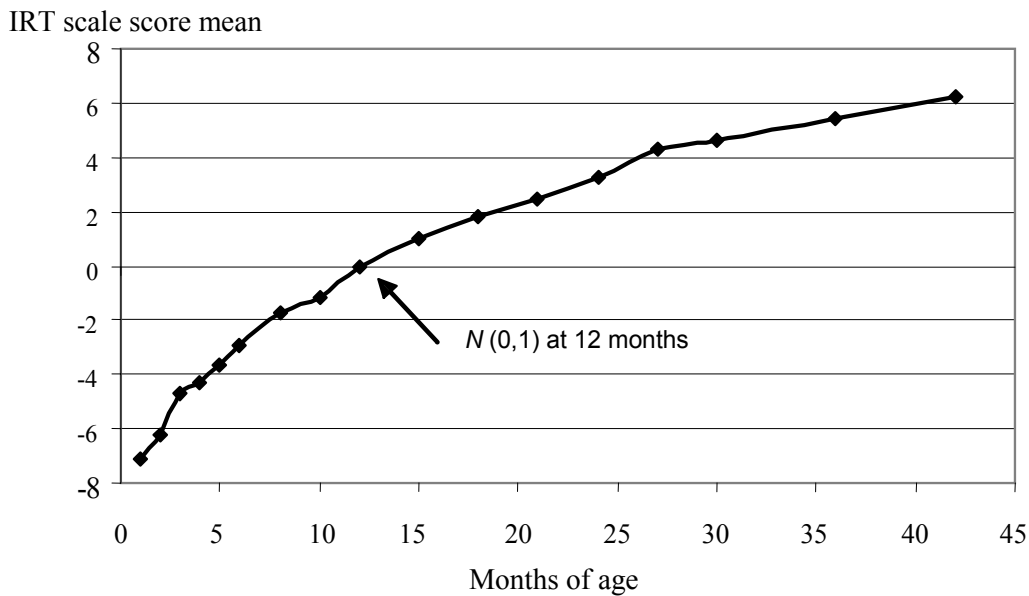


NOTE: $N(0,1)$ represents the standard normal distribution with mean = 0 and standard deviation = 1.

SOURCE: Standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

For the motor scale, population means range between $-7 < \theta < 6$ population standard deviations, as shown in figure 2-3. By working outward from the center of the scale at 12 months of age, along a sequence of age groups that are serially related by only a limited number of overlapping items in adjacent age groups, either scale is best defined toward its center, around 12 months of age. The scales tend to wobble at the extremes due to the lack of common item linkages directly relating infants at 1 and 42 months of age. The age-specific latent ability distributions have standard deviations that are nearly equal to 1, with small tendency for the variation to increase at extreme ages. Early motor development is also explosive and again slows with advancing age, similar to growth on the mental scale.

Figure 2-3. BSID-II motor scale score means by age: Item Response Theory (IRT) 2-parameter logistic item calibrations using publisher data: 1993



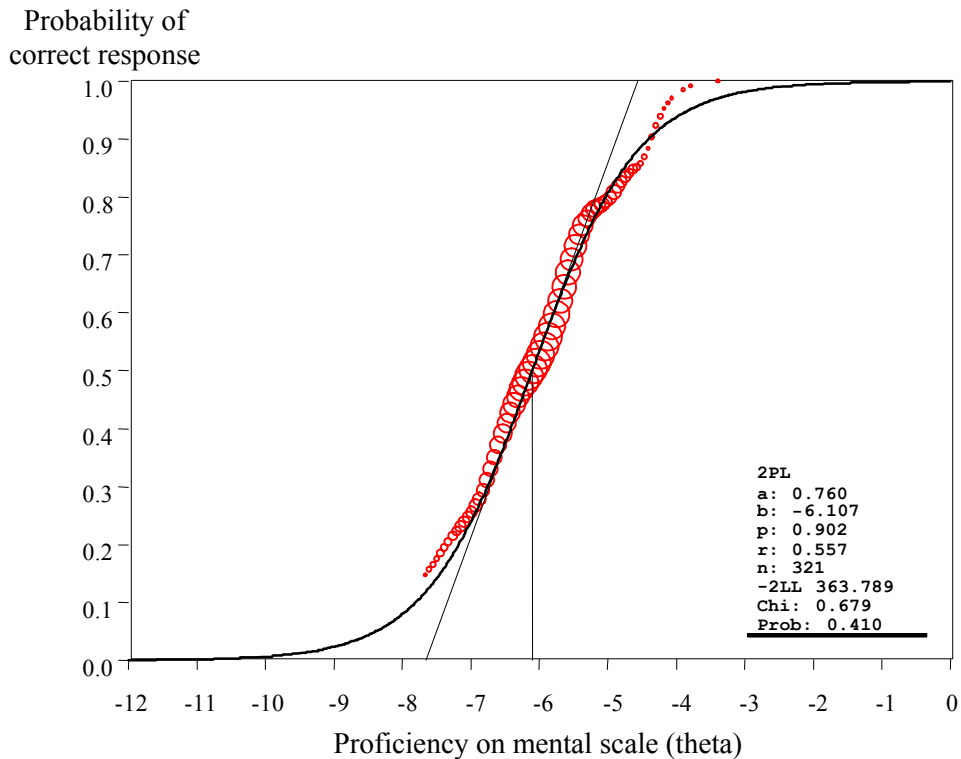
NOTE: $N(0,1)$ indicates the standard normal distribution with mean = 0 and standard deviation = 1.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Concurrent item characteristic estimation yields item calibrations similar to that shown as an example in figure 2-4. The numbering of BSID-II items is intended to reflect the item's relative difficulty. MEN028 is the 28th item among 178 mental scale items, implying that it is one of the easiest items in the BSID-II. Administration involves showing a stimulus card with two checkerboard patterns to a child and awarding credit for the item if the child gazes longer at the complex pattern. This item is recommended for children between 2 and 3 months of age. The item numbering scheme coincides with a number-right

raw score of 28 points on the publisher’s mental scale. A raw score of 28 points falls between the standardization sample means for children 2 and 3 months of age.

Figure 2-4. Item characteristic curve (ICC) for item MEN028 representing the probability of a correct response: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



NOTE: a =discrimination parameter; b =difficulty parameter; p =percentage correct; r =item-to-scale correlation; n =number of sample observations; $-2LL$ =-2 times Log Likelihood; Chi =Chi-square; $Prob$ =Significance of Chi-square; 2PL = 2-parametric logistic. Circles represent the empirical data and are proportional in size to the total number of observations at each point.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

The ICC for this item is rising most sharply opposite scale values in the vicinity of $\theta = -6.1$. Accordingly, the IRT ability parameter for this item is $b = -6.107$, reported in the box on the right in figure 2-4, and is represented by a vertical line rising to the inflection point on the ICC curve, where the probability of a correct response is exactly $P(\theta) = 0.5$. The mean age-ability relationship depicted in figure 2-2 shows that this is indeed the appropriate scale range for infants between 2 and 3 months of age.

The IRT difficulty parameter reflects the extraordinary breadth of the two BSID-II scales. The range of IRT difficulty parameters for the full set of 178 mental items is $-12.6 < b < 9.5$ and $-10.6 < b < 7.4$ for the 111 motor items. Both ranges are covered by a large number of items, implying that each scale contains many ICCs like the one shown in figure 2-4, spaced apart at short intervals averaging only 0.12 of a population standard deviation for the mental items and 0.10 for the motor items. There appear to be an abundance of items available to represent the many stages of infant development. The correlation between the IRT item difficulty parameters and item raw score rank order exceeds $r = 0.99$ for both the mental and motor item sets.

The statistics at the lower right in figure 2-4 report that the IRT discrimination parameter is $a = 0.760$, showing that this item is moderately discriminating. The a parameter is proportional to the slope of the ICC at the point of inflection, where $b = -6.107$. The slope is represented in the figure by a tangent line passing through the point of inflection, where $P(\theta) = 0.5$. Items with steeper slopes have greater discrimination and are more useful in separating examinees into different ability groups than are items that show lesser slope.

The average IRT discrimination parameter for the mental items is $a = 0.97 \pm 0.35$ and $a = 0.91 \pm 0.30$ for the motor items. Items with discrimination parameters near $a = 1$ have good discrimination. On average, the BSID-II items show good discrimination. However, there is considerable variation in item discrimination power. This suggests that the 2-PL IRT model is more appropriate for this dataset than the Rasch model, which has only an item difficulty parameter and has no provision for items that vary in discrimination. Clearly, some BSID-II items are more discriminating than others.

The circles in figure 2-4 are drawn to scale to represent the number of observations in the calibration dataset and reflect response probabilities assuming that the 2-PL response model is appropriate. When the model fits the data, the circles will align with the ICC function. Visual inspection and χ^2 statistics suggest that there are perhaps a dozen or so mental items (6 percent or 11 items in 178) that are marginally represented by the 2-PL model. Although the quality of fit also varies for motor items, it appears that, for the motor scale, virtually all of the items fit the model. With only minor shortcomings in terms of fit, all of the selected publisher items were retained in the final IRT mental and motor scales.

The information conveyed by an IRT item depends on the slope and position of the ICC. More information about an examinee's ability is obtained when the value of the a parameter is more expressive and when item difficulty b coincides with examinee ability θ . In other words, items with

considerable power of discrimination, at the appropriate level of difficulty for the examinee, convey the most information about the examinee's ability. An item may provide considerable information at one end of the ability continuum but provide no information elsewhere. Test information is a composite sum of the information provided by each of the items.

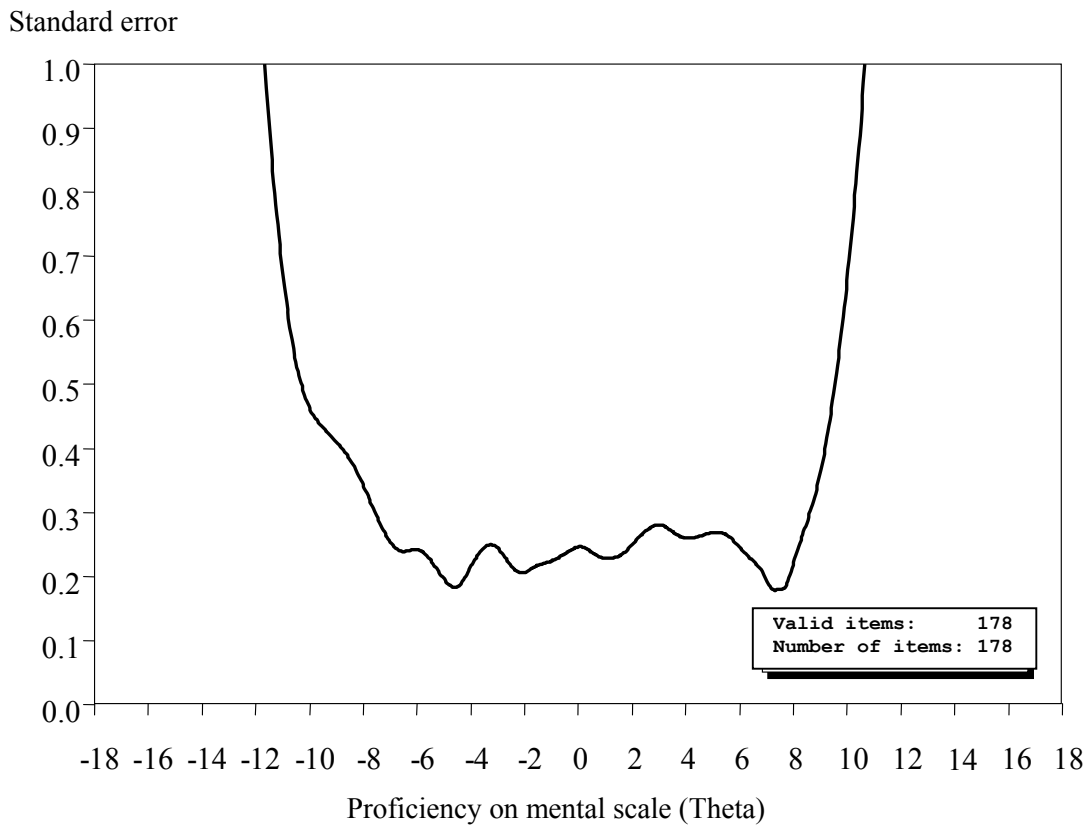
Collectively, the 178 mental and 111 motor items convey an extraordinary amount of information about children. The items are numerous, discriminate well, and are age appropriate in relation to the target population. These conditions produce tests that are both reliable and informative. High levels of information, in turn, imply that standard errors of measurement are relatively small. The standard error of measurement at different levels of ability for the IRT mental scale is shown in figure 2-5. Indeed, the standard error across most of the ability distribution is $se(\theta) < 0.3$, implying that the errors are less than one-third of a population standard deviation across virtually all of the distribution that is relevant for children between 1 and 42 months of age.

The standard error of measurement for the IRT motor scale is shown in figure 2-6. Precision of the motor scale is not as high at the extremes of the ability continuum but remains impressive across most of the ability range appropriate for infants between 1 and 42 months of age. Although information functions and standard errors are the preferred measures of test precision in IRT, a single summary index can be calculated to represent overall test reliability. Reliability represents the true score variance as a proportion of total variance and is estimated to be $r_{xx} = 0.94$ for the IRT mental scale and $r_{xx} = 0.92$ for the motor scale. These coefficients probably overstate the actual degree of test reliability since they implicitly assume that the full set of items will be used. Nevertheless, they appear to be consistent with publisher documentation reporting high levels of reliability for conventional BSID-II scales, with KR-20 (Kuder and Richardson 1937⁵) coefficients of internal consistency averaging $r_{xx} = 0.88$ for the mental scale and $r_{xx} = 0.84$ for the motor scale across all age groups.⁶

⁵ The Kuder-Richardson 20 statistic measures test reliability of inter-item consistency. A higher value indicates a strong relationship between items on the test.

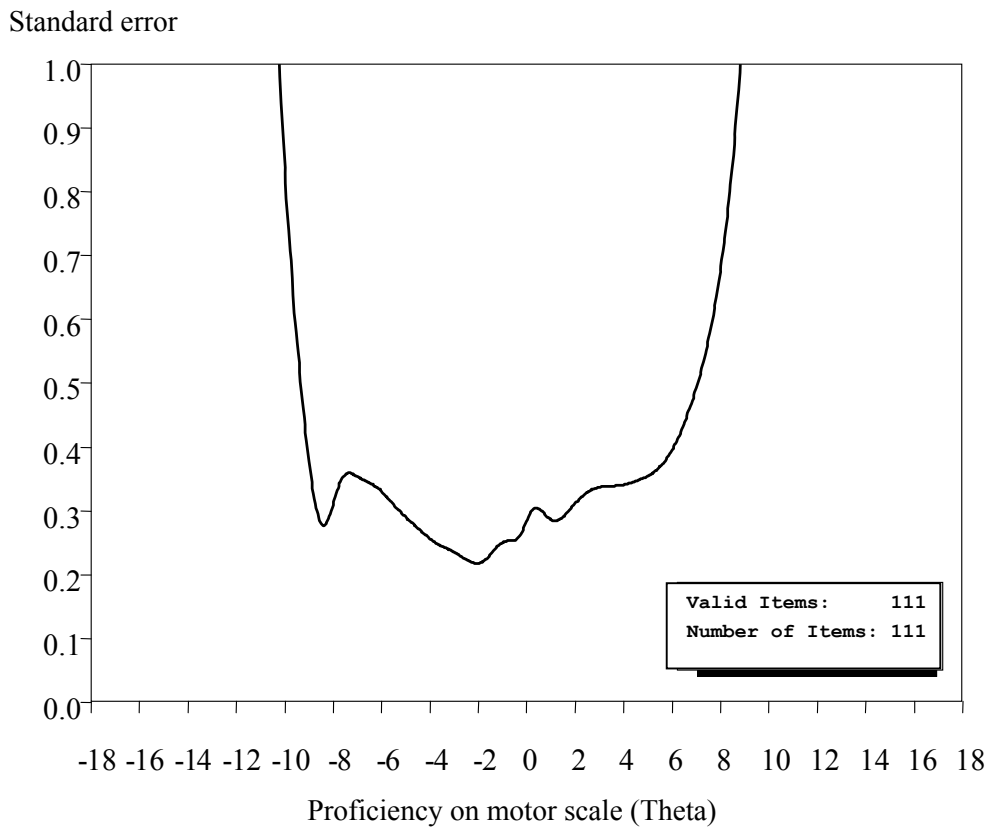
⁶ These coefficients are IRT equivalents of KR-20 coefficients. Although similar to coefficient alpha, the more general symbol for reliability, r_{xx} , is used.

Figure 2-5. Standard error of measurement for the mental scale: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



SOURCE: Standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Figure 2-6. Standard error of measurement for the motor scale: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993

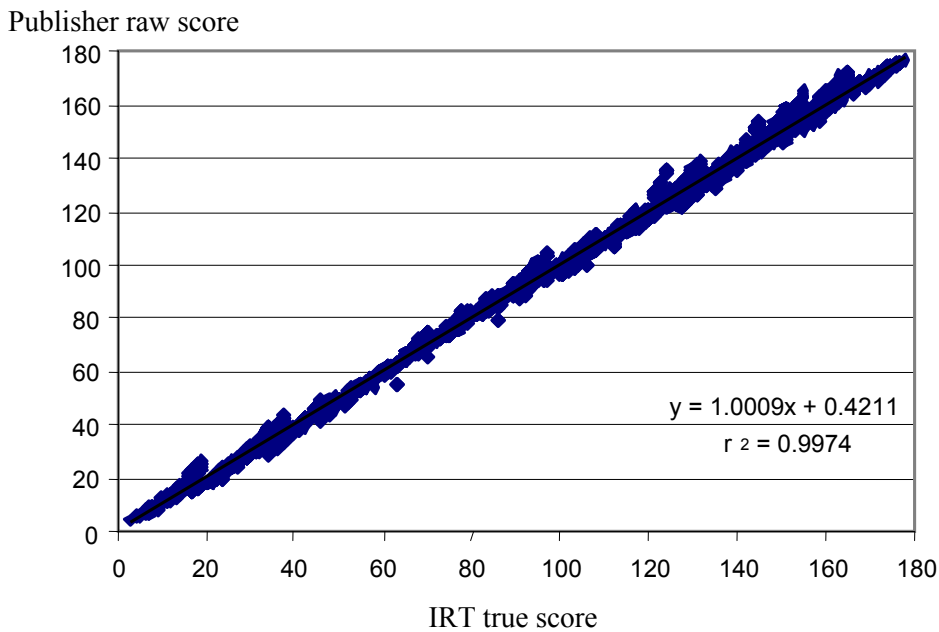


SOURCE: Standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

The objective of testing is to assign a score to an individual examinee that reflects the level of attainment of a skill. One approach to scoring is to give a point for each correct response and present the test outcome as an item-correct raw score. Indeed, this is the origin of the number-right raw score metric used by the publisher to provide national norms for the BSID-II scales. The only difficulty with this approach is that, by adding items to or subtracting items from the test, the raw score metric will change. Obtaining 14 correct responses out of 20 is different from obtaining 14 right out of 50. A method must be found to permit item substitution and deletions without altering the scale metric used to express test results. IRT has been developed to enable this flexibility. However, first it must be shown that IRT ability estimates θ can be reported, using publisher raw score metric.

IRT item calibrations enable the prediction of number-right raw scores. In IRT, the functional equivalent of the number-right raw score is the IRT true score. The IRT true score is the expected number of correct responses, expressed in the same metric as the number-right raw score. This is the sum of item probabilities $P_j(\theta)$ across all items j at a specific level of ability θ . $\xi = \sum_{j=1}^n P_j(\theta)$. As a final check on the quality of item calibrations, figure 2-7 shows the relationship between IRT true scores and raw scores for the mental scale, using observations in publisher data.

Figure 2-7. Relationship between Item Response Theory (IRT) true score and publisher raw score for the mental scale: IRT 2-parameter logistic item calibrations using publisher data: 1993

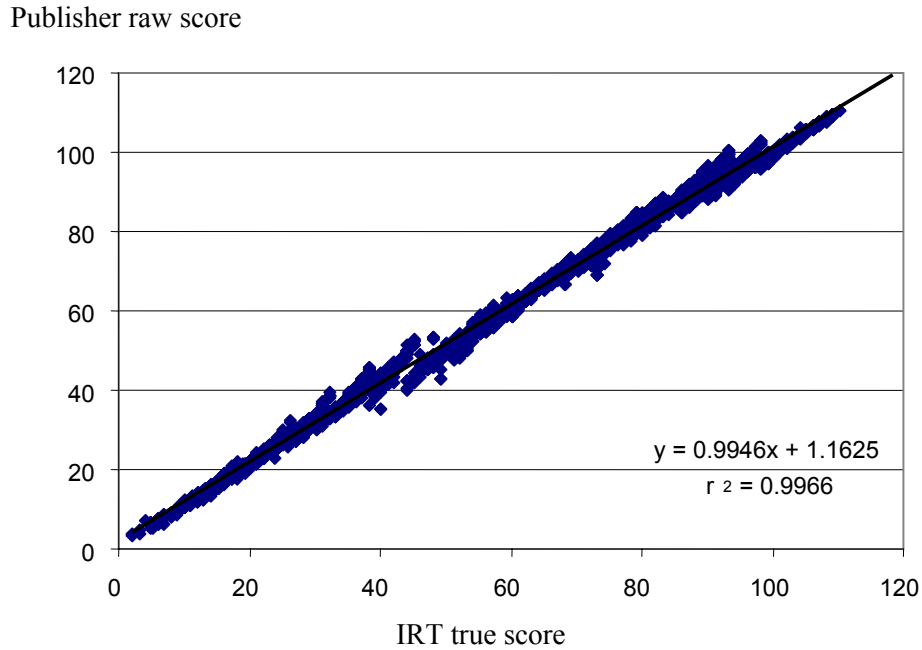


SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

The linear relationship between raw scores and IRT true scores has its origin near zero ($a = 0.421$ on a 178-point scale), a slope coefficient that is almost exactly one (to three decimal places $b = 1.000$), and a coefficient of determination that is almost unity ($r^2 = 0.997$). Figure 2-8 shows essentially identical results for the motor scale, with an origin near zero ($a = 1.1625$ on a 111-point scale), a slope coefficient that is almost exactly one ($b = 0.995$), and a coefficient of determination that is again almost unity ($r^2 = 0.995$). These relationships show that it is possible to express IRT ability estimates in

raw score metric. This, in turn, is the key to reporting standardized scores that allow direct comparisons among infants of different ages.⁷

Figure 2-8. Relationship between Item Response Theory (IRT) true score and publisher raw score for the motor scale: IRT 2-parameter logistic item calibrations using publisher data: 1993



SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

2.1.7 Constructing Item Sets for the 18-Month BSF-R Mental and Motor Scales

Constructing the item sets for the BSF-R mental and motor scales was a multistep process that involved selecting the items to be included in each scale, then developing the decision rules that would be used to route children from the core set of items to the basal set or ceiling sets, if necessary, followed by examination of the projected reliability that could ideally be attained (based on the BSID-II standardization data).

⁷ Standardized scores are reported by the publisher as *development index scores*. In the ECLS-B, standardized scores are called *T-scores*.

Selecting Items for the BSF-R Mental Scale

Once the 178 mental and 111 motor items were calibrated using publisher data, it became possible to predict how individuals will respond to items before any test is taken into the field. Item parameters define an item response function representing the probability of a correct response by any examinee. This can be used to make predictions about how people will behave in the real world. An almost endless variety of hypothetical tests can be constructed from these same item pools, and their technical properties can be examined before any such test is selected for production or goes into the field. Alternative tests can be tailored to any ability level and adapted as needed to provide levels of reliability.

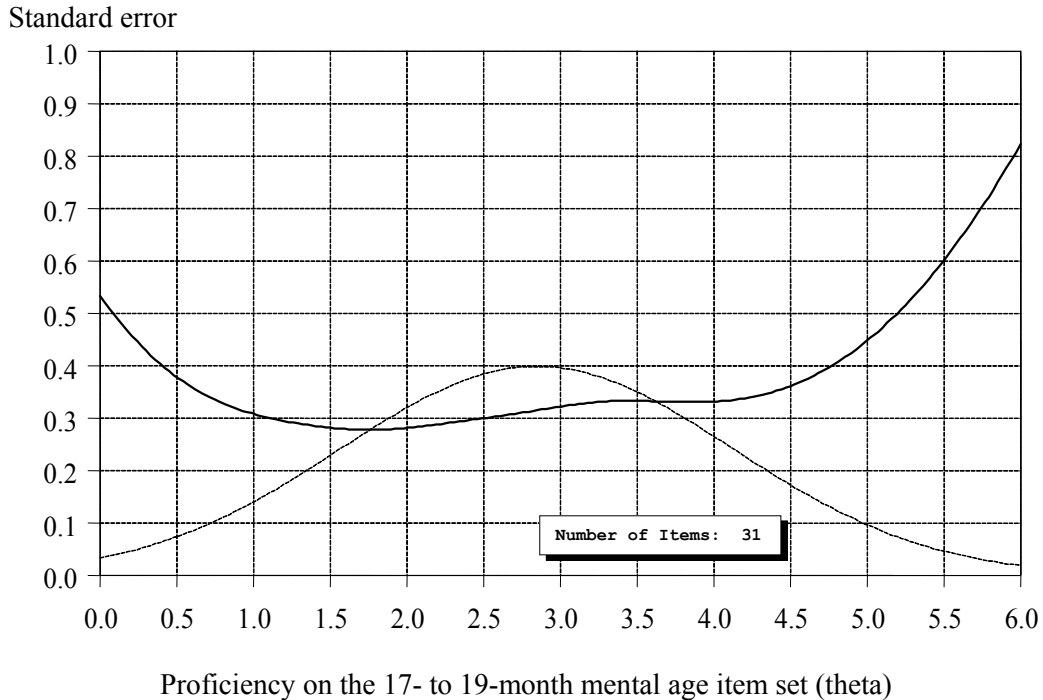
In order to select reduced item sets for the 18-month BSF-R, the technical properties (i.e., difficulty and discrimination parameters) of items in the 18-month⁸ age item set recommended by the publisher were examined. There are 100 observations for this age group in the publisher standardization sample.

Selecting Items for the 18-Month BSF-R Mental Scale

Figure 2-9 shows the respective target population ability distribution superimposed on a graph of the standard error of measurement $se(\theta)$ obtained for the 31 items in the publisher-recommended 18-month age item set for children 17 to 19 months of age. It represents the projected standard error of measurement that would probably be obtained if the entire age set of 31 items were administered. For reference purposes, the 18-month frequency distribution appears in the background and is represented by a dashed line. Approximately 68 percent of the population falls within $\mu \pm \sigma$.

⁸ The item age set for 18-months actually ranges from 17 to 19 months but is referred to as the 18-month age set for convenience.

Figure 2-9. Standard error of measurement by proficiency level for the publisher-recommended 17- to 19-month BSID-II mental age item set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



NOTE: The publisher does not provide data for all months of age in the calibration dataset but does provide recommended items for different months of age. The normal curve (dashed line) represents the projected latent distribution of the 17- to 19-month population and is included for illustration purposes.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID II), The Psychological Corporation, 1993.

The standard error depicted in the figure shows that the 31 items in the age item set recommended by the publisher afford considerable measurement precision for 18-month-old children within the limits of $\mu \pm \sigma$. Moving outward from the mean, growth in the standard error of measurement accelerates, and beyond $\mu \pm \sigma$ the standard error increases very rapidly. For some purposes, the error $se(\theta) > 0.5$ outside roughly $\theta \pm 1.5\sigma$ might be considered excessive. This is why the publisher recommends testing the limits of each child's ability with the recommended age item set, followed by the administration of basal and ceiling item sets as required. In this event, all of the items in the adjacent item set or sets are to be administered.

For use in the ECLS-B, the BSF-R was designed to reduce administration time without compromising the quality of the child development data that are collected. The BSF-R was also designed to replicate results obtained with the BSID-II as closely as possible. This was accomplished by selecting

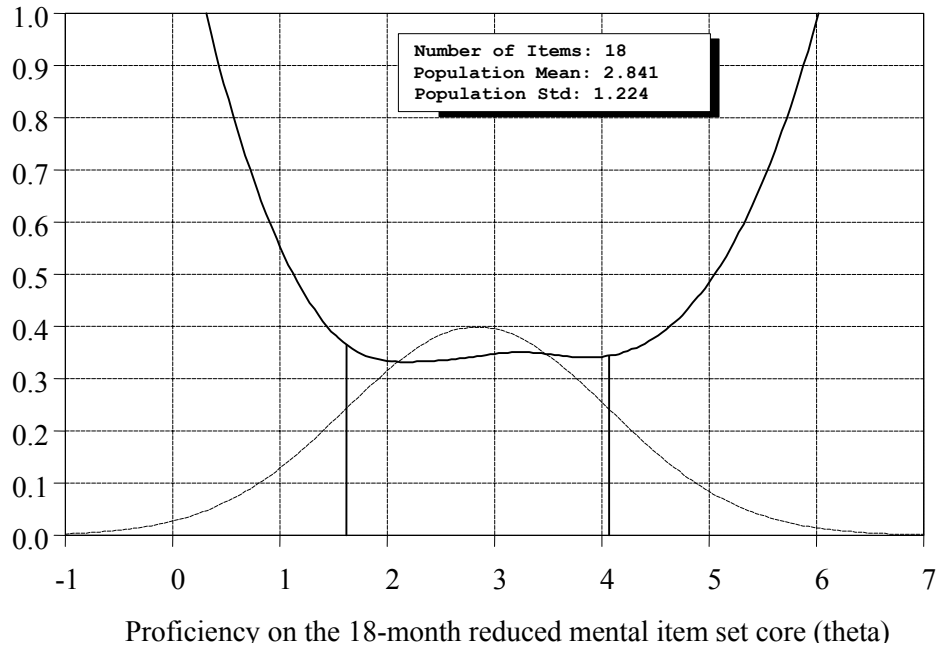
smaller item sets from the BSID-II item pool and using an adaptive testing strategy. Assessment workgroup members advised that, ideally, the BSF-R should yield standard errors of measurement in the vicinity of $se(\theta) = 0.4$ across the target population ability distribution, extending well out into the tails. This corresponds with a reliability coefficient of approximately $r_{xx} = 0.8$.

The selection of reduced item sets for BSF-R began by examining the most highly discriminating items available in the range of difficulty appropriate for 18-month-old children. For the core item set, this is approximately $\mu \pm \sigma$. Item selection began by considering items with IRT difficulty parameter values that extend slightly beyond the range of $\mu \pm \sigma$. Within this general range of difficulty, priority of selection was given to the most discriminating items, those where item discrimination parameter values exceed $a > 0.9$. Consideration was given to item content coverage and ease of administration before selecting a final item set.

Based on these criteria, reduced core item sets were constructed with desirable measurement properties appropriate for children in an age-specific target population. The approach used in the BSF-R is illustrated beginning with the standard errors for the 18-month mental core item set presented in figure 2-10. A set of 18 items satisfied all of the above criteria and was used to construct a hypothetical mental scale core item set based on the new reduced set of items. Items calibrated with publisher data can now be used to estimate the new core item set's standard error of measurement across the full range of ability. The new scale is not quite as precise as the 31-item scale based on the publisher's recommended age item set. However, the reduced item set affords standard errors that meet or exceed the objective $se(\theta) = 0.4$ over the range $\mu \pm \sigma$.

Figure 2-10. Standard error of measurement by proficiency level for the 18-month BSF-R mental core item set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993

Standard error



NOTE: The solid lines descending from the curve to the x-axis indicate 1 standard deviation above and 1 standard deviation below the mean. Std = standard deviation. The normal curve (dashed line) represents the projected latent distribution of the 18-month population and is included for illustration purposes.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

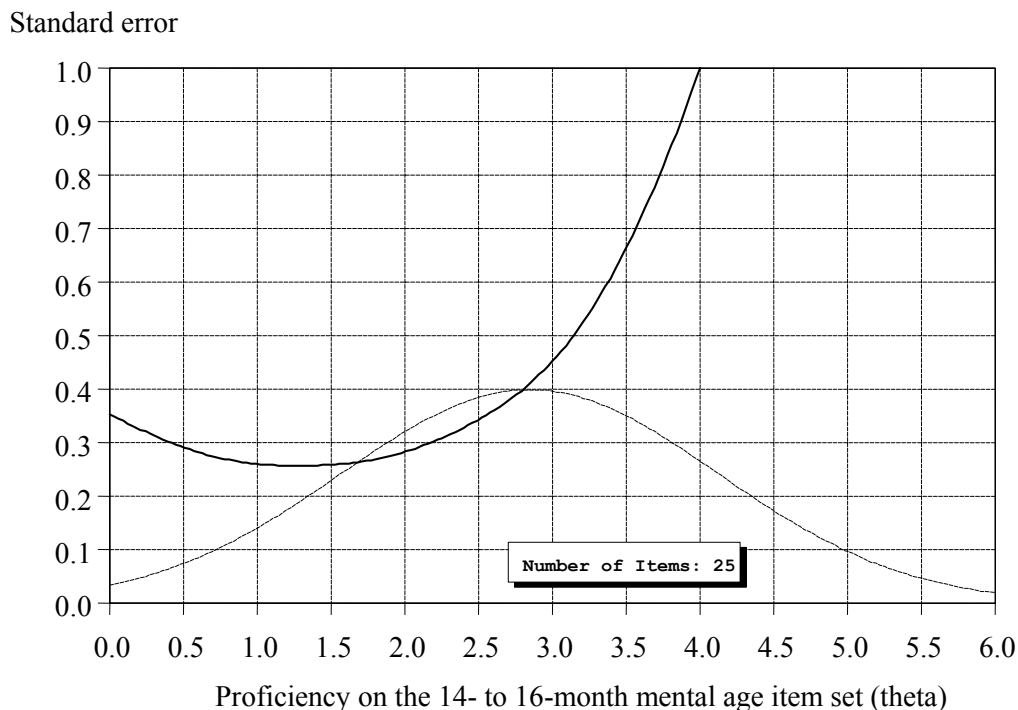
The new mental core item set yields satisfactory precision across the central part of the target population's latent distribution, between $\mu \pm \sigma$, where approximately 68 percent of children are to be found, meaning that the assessment of many children will require the administration of no more than 18 items. Results obtained with the core items would be sufficiently precise to produce ability estimates within an acceptable margin of error in the middle of the ability distribution. Depending on the outcome obtained with this initial core set, basal, or ceiling items would then be administered only to those children that require them.

Outside the range $\mu \pm \sigma$ (i.e., more than 1 standard deviation from the mean in either direction), appropriate basal and ceiling items would have to be administered so that the objective $se(\theta) = 0.4$ will be satisfied at the tails of the distribution. In addition, a decision rule governing the application of basal and ceiling items, based on results obtained with the initial core set, needs to be

established. This strategy for adaptive testing would yield appropriate measures for all of the children, including those with exceptional levels of ability, in the age group while still reducing the burden of fieldwork.

The BSID-II item pool was again consulted to find items for the tails of the ability distribution. Successive age item sets were examined, and IRT analyses found the 14- to 16-month set of 25 items (figure 2-11) to be a likely source of highly discriminating basal items, appropriate for the 18-month population scoring below $\mu - \sigma$. These items actually ranged in difficulty from 12 to 22 months so that all were also within the 17- to 19-month age set. IRT difficulty and discrimination parameters b and a were examined together with considerations of item coverage and ease of administration before proceeding with item selection. On this basis, a reduced mental basal set of eight items was selected, to be administered only as a complement to the BSF-R mental core item set. Consequently, it was not necessary to examine the technical properties of a hypothetical scale comprising basal items alone.

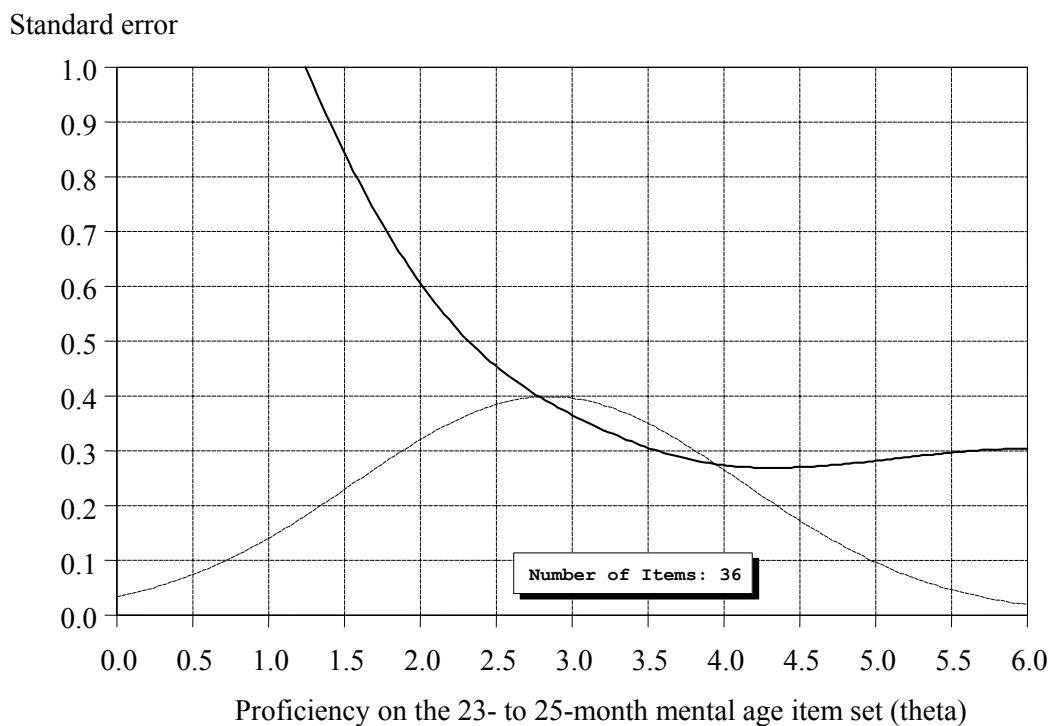
Figure 2-11. Standard error of measurement by proficiency level for the BSID-II mental scale 14- to 16-month age item set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



NOTE: The publisher does not provide data for all months of age in the calibration dataset but does provide recommended items for different months of age. The normal curve (dashed line) represents the projected latent distribution of the 14- to 16-month population and is included for illustration purposes. SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

At the upper end of the ability distribution, items needed to be found for a reduced mental ceiling item set (figure 2-12). IRT analyses demonstrated that the 23- to 25-month mental age item set of 36 items was a good source of ceiling items at the required level of difficulty of $\mu + \sigma$. These items ranged in difficulty from 17 to 42 months of age. IRT difficulty and discrimination parameters b and a were examined together with considerations of item coverage and ease of administration before proceeding with item selection. On this basis, a reduced mental ceiling set of nine items was selected, to be administered only as a complement to the BSF-R mental core item set when necessary. Consequently, it was not necessary to examine the technical properties of a hypothetical scale comprising basal items alone.

Figure 2-12. Standard error of measurement by proficiency level for the BSID-II mental scale 23- to 25-month age item set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



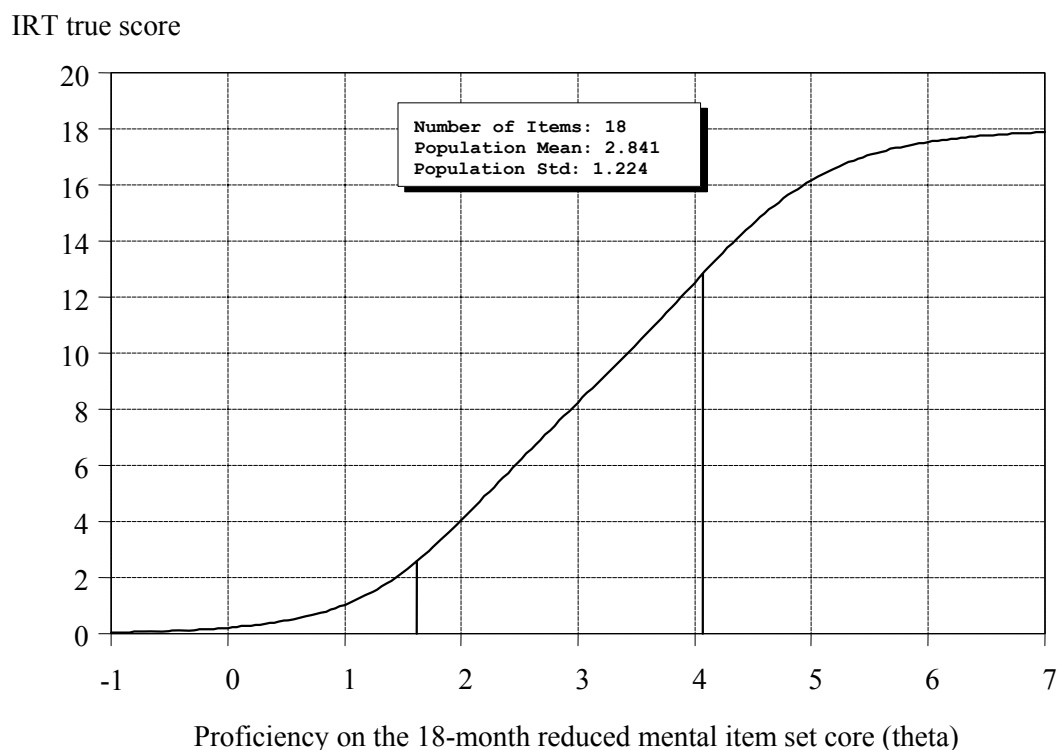
NOTE: The publisher does not provide data for all months of age in the calibration dataset but does provide recommended items for different months of age. The normal curve (dashed line) represents the projected latent distribution of the 23- to 25-month population and is included for illustration purposes.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Using IRT to Develop Basal and Ceiling Decision Rules for the 18-Month BSF-R Mental Scale

For the adaptive testing strategy to work properly, basal and ceiling decision rules needed to be devised that were simple enough so that they could be easily followed in the field. A straightforward rule based only on counting the number of correct responses (i.e., the raw score) would be easier in the field than the rules used in the full BSID-II, which involved summing the number of correct scores (to route to the basal set) and the number of incorrect scores (to route to the ceiling set). The functional equivalent of the raw score in IRT is the expected number-right or IRT true score. This is simply the sum of the probabilities of a correct response across all items at a given level of ability. The IRT true score for the 18-month mental reduced core item set is shown in figure 2-13.

Figure 2-13. Establishing basal and ceiling rules for the 18-month BSF-R mental core item set using true scale scores: Item Response Theory (IRT) 2-parameter logistic item calibrations using publisher data: 1993



NOTE: The solid lines descending from the curve to the x-axis indicate 1 standard deviation above and 1 standard deviation below the mean. Std = standard deviation.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

The IRT true score for the reduced core item set is zero at extremely low levels of ability, rises rapidly across the central range of the latent distribution, and approaches the total number of items in the core item set at high levels of ability. Measurement precision is highest across the range where the expected true score is rising most rapidly. This coincides with core item difficulty levels located in the range of $\mu \pm \sigma$, which is again delimited by a pair of vertical lines in the figure. Rules were defined at the limits of this range so that decisions could be made to determine whether either the basal or ceiling item set needed to be administered.

Reading the true score value opposite each of the vertical lines at the point where they join the curve provides an estimate of the number-right score at each of these limits. The values are approximately 3 at the lower end and 13 at the high end of this range. The conservative decision rule that was actually defined for the BSF-R mental scale at the low end is that a score of 0 to 4 points on the core item set requires administration of the basal item set. At the high end, the rule is that a score of 12 to 18 points on the core item set requires administration of the ceiling item set.

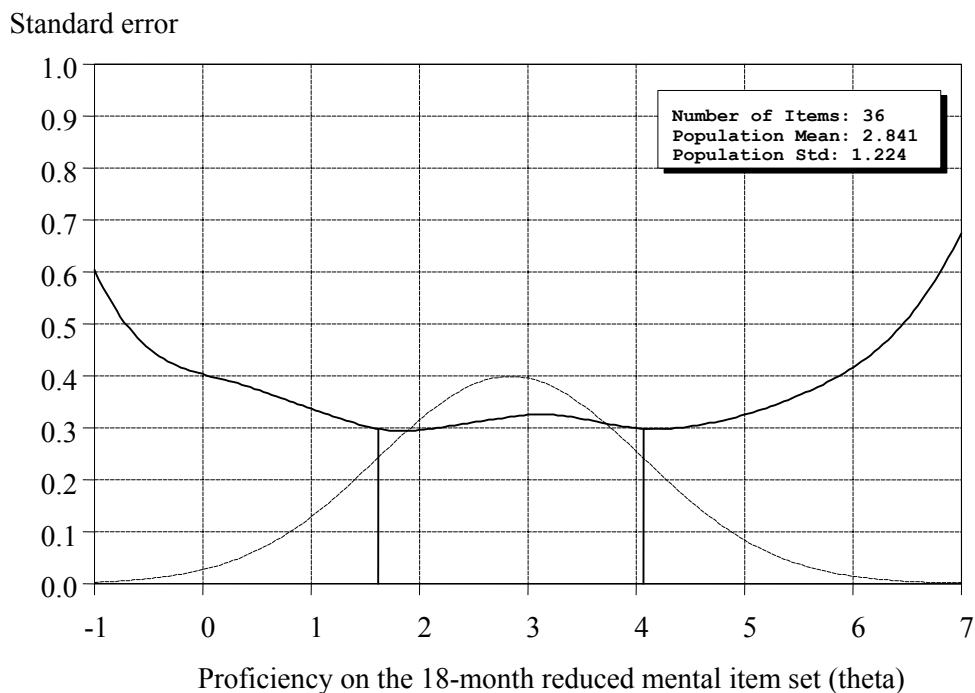
The 8 basal items, 18 core items, and 10 ceiling items contribute a total of 36 items to the 18-month BSF-R mental scale. A child would never be administered all of these items. Neither would the basal or ceiling items be administered by themselves but rather only after first administering the core item set. Consequently, a child can be administered either 18, 26, or 28 items, depending on whether the core items are sufficient or whether the basal or ceiling items are also required. Approximately 68 percent of the target population will receive only the 18 core items. Another 32 percent will also be administered either the basal set or the ceiling set. It may help to think of it as a weighted average based on the expectation that 68 percent receive only 18 items, whereas another 16 percent receive 26 items and the remaining 16 percent receive 28 items, so that on average across the entire sample, 21 items are administered to each child. Consequently, the expected average is $(68 \text{ percent} \times 18) + (16 \text{ percent} \times 26) + (16 \text{ percent} \times 28) = 21$ mental items administered, on average.

Projected Standard Error of the 18-Month BSF-R Mental Scale

Figure 2-14 shows the standard error for the 18-month BSF-R mental scale of 36 items (i.e., all basal, core, and ceiling items), based on item calibrations obtained with publisher data. Although the figure is based on all 36 items, it is at least approximately correct for the core, basal, and ceiling item combinations that were found in practice. This is because the basal items have relatively little impact on

standard error at the middle of the distribution and virtually no impact at the high end of the distribution. Ceiling items have little impact on standard errors at the middle of the distribution and virtually no impact at the low end of the distribution. Conceivably, subjects who are administered only the core item set will have somewhat larger errors than those depicted in the figure if their abilities lie at the limits of $\mu \pm \sigma$. IRT item calibrations based on ECLS-B data will yield somewhat different standard errors than those from publisher data that are depicted in figure 2-14.

Figure 2-14. Projected standard error of measurement by proficiency level for the 18-month BSF-R mental scale: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



NOTE: The solid lines descending from the curve to the x-axis indicate 1 standard deviation above and 1 standard deviation below the mean. Std = Standard deviation. The normal curve (dashed line) represents the projected latent distribution of the 18-month population and is included for illustration purposes.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Items in the 18-Month BSF-R Mental Scale

Exhibit 2-2 lists the items in the 18-month BSF-R mental scale, by basal, core, and ceiling sets.

Exhibit 2-2. 18-month BSF-R mental scale items, by item set: 2003–04

| BSID-II item number | Item description |
|---------------------|--------------------------------------|
| Basal items | |
| MEN086 | Puts three cubes in cup |
| MEN089 | Puts six beads in box |
| MEN091 | Scribbles spontaneously |
| MEN094 | Imitates word |
| MEN095 | Puts nine cubes in cup |
| MEN097 | Builds tower of two cubes |
| MEN102 | Retrieves toy |
| MEN100 | Uses two different words |
| Core items | |
| MEN099 | Points to two pictures |
| MEN106 | Uses word(s) to make wants known |
| MEN107 | Follows directions (doll) |
| MEN108 | Points to three of doll's body parts |
| MEN109 | Names one picture |
| MEN110 | Names one object |
| MEN111 | Combines word and gesture |
| MEN113 | Says eight different words |
| MEN114 | Uses a two-word utterance |
| MEN121 | Uses pronoun(s) |
| MEN122 | Points to five pictures |
| MEN123 | Builds tower of six cubes |
| MEN124 | Discriminates book, cube and key |
| MEN125 | Matches pictures |
| MEN126 | Names three objects |
| MEN127 | Uses a three-word sentence |
| MEN128 | Matches three colors |
| MEN131 | Attends to story |
| Ceiling items | |
| MEN129 | Makes a contingent utterance |
| MEN133 | Names five pictures |
| MEN134 | Displays verbal comprehension |
| MEN135 | Builds tower of eight cubes |
| MEN136 | Poses question(s) |
| MEN137 | Matches four colors |
| MEN140 | Understands 2 prepositions |
| MEN141 | Understands concept of one |
| MEN142 | Talks in response to picture book |
| MEN144 | Discriminates pictures |

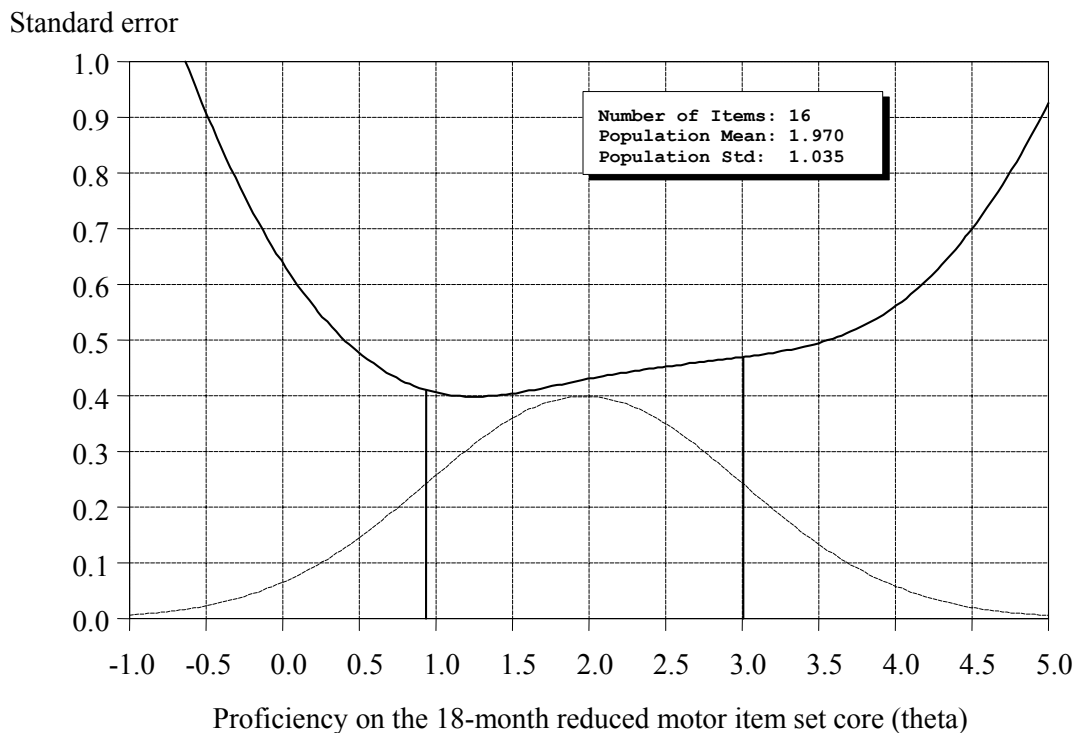
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 2-year data collection, 2003–04.

Selecting Items for the 18-Month BSF-R Motor Scale

Construction of the 18-month BSF-R motor scale followed the same procedures summarized above for the mental scale; appropriate age item sets were identified for the core item set; feasible items were selected for the basal and ceiling sets; and the standard error of the measure (*SEm*) was examined using the publisher standardization dataset.

The publisher-recommended age item set was identified as the 17- to 19-month set. These items were reviewed and tested. Items were eliminated that required complicated materials (e.g., a set of steps built to standard specifications), that were too difficult to administer, or that were too subjective to score. The standard error of the candidate items for the motor scale core set is presented in figure 2-15.

Figure 2-15. Standard error of measurement by proficiency level for the 18-month BSF-R motor scale core set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



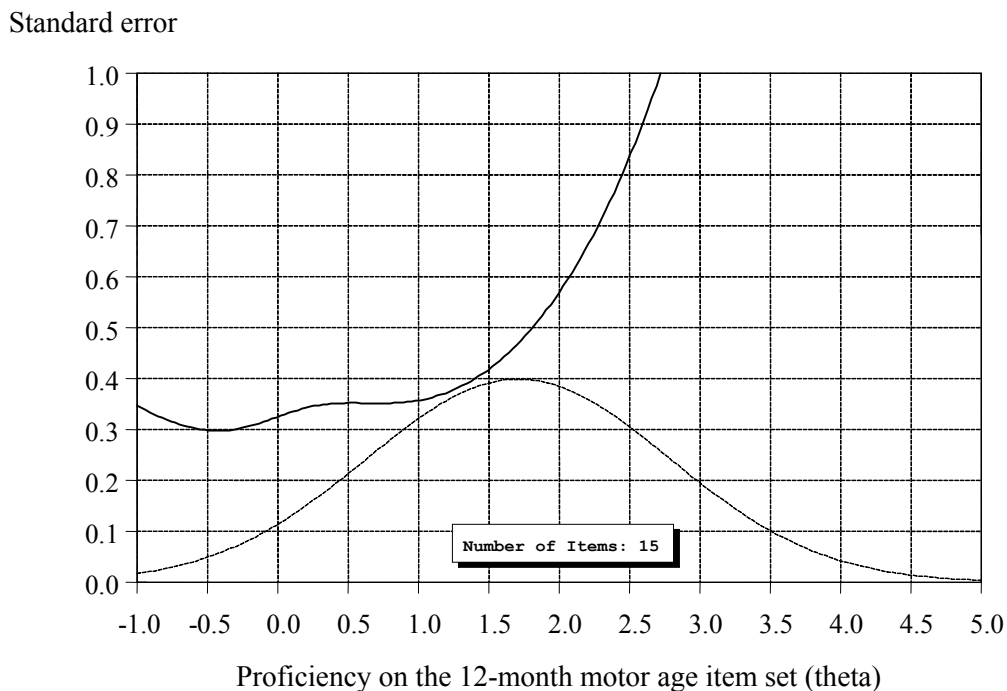
NOTE: The solid lines descending from the curve to the x-axis indicate 1 standard deviation above and 1 standard deviation below the mean. Std = standard deviation. The normal curve (dashed line) represents the projected latent distribution of the 18-month population and is included for illustration purposes.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

This figure shows that the standard error is a bit high, more so at the upper end of the core set, where it approaches 0.5, than at the lower end, where the standard error is nearer to 0.4. However, this was remedied by careful construction of the basal and ceiling sets.

IRT analysis determined that the 12-month age set was a good source for basal items, as shown in figure 2-16.

Figure 2-16. Standard error of measurement by proficiency level for the BSID-II 12-month motor age item set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993

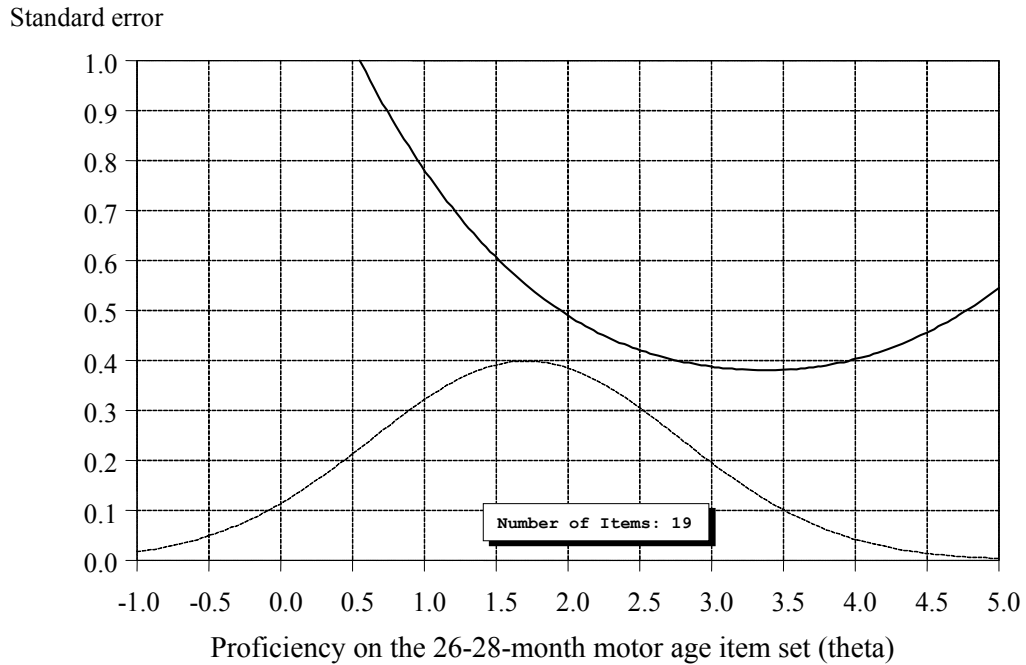


NOTE: The publisher does not provide data for all months of age in the calibration dataset but does provide recommended items for different months of age. The normal curve (dashed line) represents the projected latent distribution of the 12-month population and is included for illustration purposes.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

IRT analysis determined that the 26- to 28-month motor age item was a good source for ceiling items, as shown in figure 2-17, although some items were also taken from the 29- to 31-month age set to complete the set of ceiling items.

Figure 2-17. Standard error of measurement by proficiency level for the BSID-II motor scale 26- to 28-month age item set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



NOTE: The publisher does not provide data for all months of age in the calibration dataset but does provide recommended items for different months of age. The normal curve (dashed line) represents the projected latent distribution of the 26- to 28-month population and is included for illustration purposes.

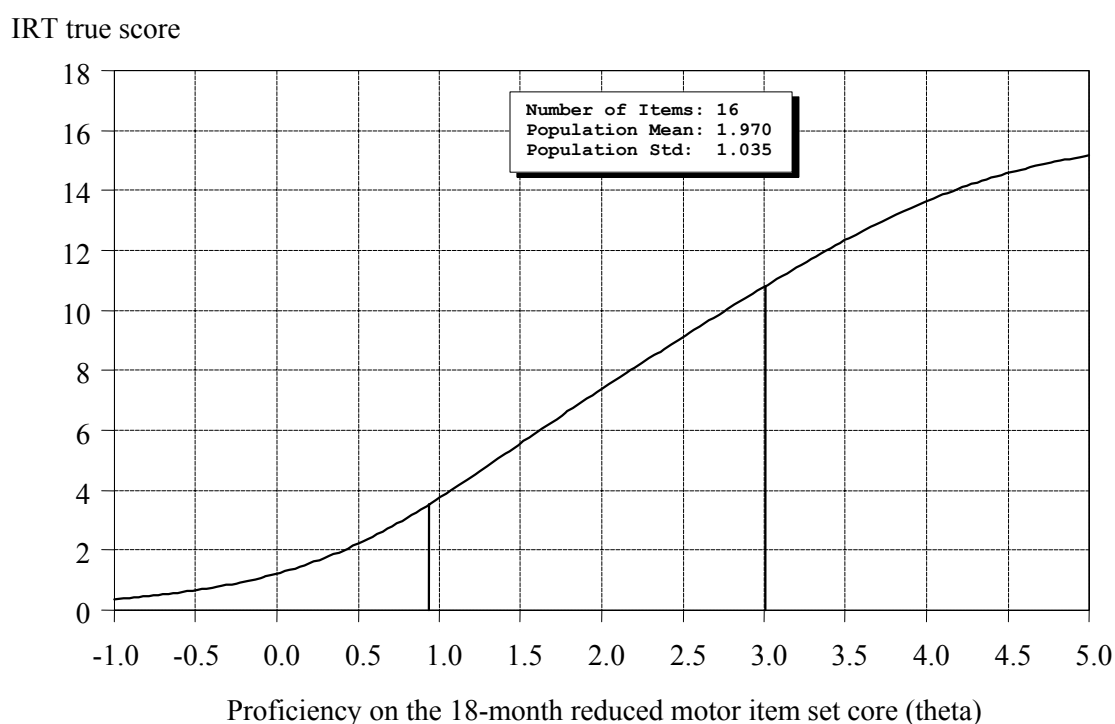
SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Based on the difficulty parameter and the discrimination parameter of the items in this age set, appropriate items were selected for the core, basal, and ceiling sets and were pilot-tested for feasibility. Items that were not feasible operationally were eliminated.

Using IRT to Develop Basal and Ceiling Decision Rules for the 18-Month BSF-R Motor Scale

A final step in developing the 18-Month BSF-R motor scale was to determine the appropriate rules for routing children to the motor basal item set or to the motor ceiling item set. IRT analyses provided the necessary information to develop these rules, as reflected in figure 2-18.

Figure 2-18. Establishing basal and ceiling rules for the 18-month BSF-R motor core item set using true scale scores: Item Response Theory (IRT) 2-parameter logistic item calibrations using publisher data: 1993



NOTE: The solid lines descending from the curve to the x-axis indicate 1 standard deviation above and 1 standard deviation below the mean. Std = standard deviation.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Reading the true score value opposite each of the vertical lines at the point where they join the curve provides an estimate of the number-right score at each of these limits. The values are slightly more than 3 at the lower end and approximately 11 at the high end of this range. The conservative decision rule that was actually defined for the BSF-R motor scale at the low end is that a score of 0 to 4 points on the core item set requires administration of the basal item set. At the high end, the rule is that a score of 12 or more points on the core item set requires administration of the ceiling item set.

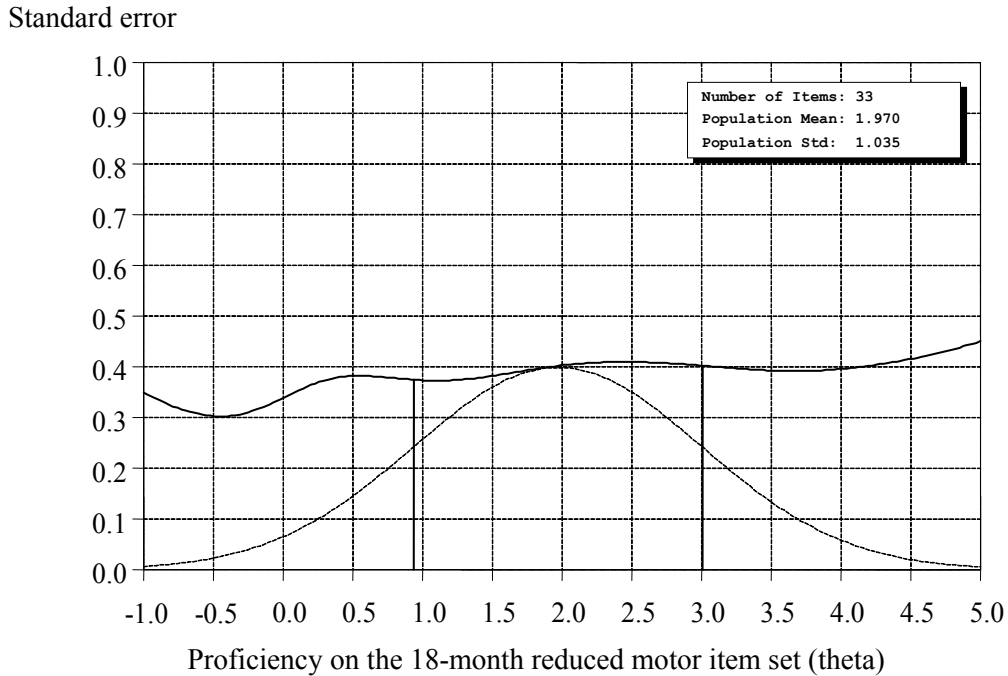
As with the mental scale, a child would never be administered all of these items. Neither would the basal or ceiling items be administered by themselves, but only after first administering the core item set. Consequently, a child can be administered either the core set of items alone, the core plus basal item set, or the core plus ceiling item set, depending on whether the core items are sufficient or whether the basal or ceiling items are also required. Approximately 68 percent of the target population will receive only the core items. Another 32 percent will also be administered either the basal set or the ceiling set. It may help to think of it as a weighted average based on the expectation that 68 percent only receive 17 items whereas another 16 percent receive 24 items and the remaining 16 percent receive 27 items, so that on average across the entire sample, 20 items are administered to each child. Consequently, the expected average is $(68 \text{ percent} \times 17) + (16 \text{ percent} \times 24) + (16 \text{ percent} \times 27) = 19.72$ motor items administered on average.

Projected Standard Error of the 18-Month BSF-R Motor Scale

Figure 2-19 shows the projected standard error for the 18-month BSF-R motor scale of 33 items⁹ (i.e., all basal, core, and ceiling items), based on item calibrations obtained with publisher data. IRT item calibrations based on ECLS-B data will yield somewhat different standard errors than those from publisher data. Nevertheless, figure 2-19 is at least approximately correct for the BSF-R core, basal, and ceiling item combinations that will be found in practice. Conceivably, subjects administered only the core item set may have somewhat larger errors than those depicted in the figure if their abilities lie at the limits of $\mu \pm \sigma$.

⁹ These analyses are based on 33 items. To bring down the standard error of mean at the higher end of the distribution 1 item was added because it was feasible to administer and had desirable psychometric properties. Therefore, the final version had 34 items.

Figure 2-19. Projected standard error of measurement by proficiency level for the 18-month BSF-R motor scale: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



NOTE: Figure 2-19 shows that the standard error in the highest tail of the distribution was slightly above the target 0.4. Therefore, items with ability parameters in the range of 4.25 to 4.75 were examined for feasibility of administration and the one most feasible was added. Therefore, the figure is based on 33 items, but there are 34 items included in the 18-month BSF-R motor scale. The solid lines descending from the curve to the x-axis indicate 1 standard deviation above and 1 standard deviation below the mean. Std = standard deviation. The normal curve (dashed line) represents the projected latent distribution of the 18-month population and is included for illustration purposes.

SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Items in the 18-Month BSF-R Motor Scale

Exhibit 2-3 is the final list of items selected for the 18-month BSF-R motor scale, as implemented in the spring 2001 field test, grouped by core, basal, and ceiling sets.

Exhibit 2-3. 18-month BSF-R motor items, grouped by core, basal, and ceiling sets: 2001

| BSID-II item number | Item description |
|---------------------|---|
| Basal items | |
| MOT058 | Grasps pencil at farthest end |
| MOT059 | Stands up I |
| MOT060 | Walks with help |
| MOT061 | Stands alone |
| MOT062 | Walks alone |
| MOT063 | Walks alone with good coordination |
| MOT064 | Throws ball |
| Core items | |
| MOT067 | Walks backward |
| MOT068 | Stands up II |
| MOT070 | Grasps pencil at middle |
| MOT072 | Stands on right foot with help |
| MOT073 | Stands on left foot with help |
| MOT074 | Uses pads of fingertips to grasp pencil |
| MOT075 | Uses hand to hold paper in place |
| MOT077 | Runs with good coordination |
| MOT078 | Jumps off floor (both feet) |
| MOT082 | Stands alone on right foot |
| MOT083 | Stands alone on left foot |
| MOT084 | Walks forward on line |
| MOT085 | Walks backward close to line |
| MOT086 | Swings leg to kick ball |
| MOT087 | Jumps distance of 4 inches |
| MOT089 | Walks on tiptoe for four steps |
| MOT090 | Grasps pencil at nearest end |
| Ceiling items | |
| MOT088 | Laces three beads |
| MOT091 | Imitates hand movements |
| MOT093 | Manipulates pencil in hand |
| MOT094 | Stands up III |
| MOT096 | Copies circle |
| MOT098 | Imitates postures |
| MOT099 | Walks on tiptoe for 9 feet |
| MOT100 | Stops from a full run |
| MOT101 | Buttons one button |
| MOT103 | Stands alone on right foot 4 seconds |

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 18-month field test, 2001.

2.1.8 Design of Administration Booklet, Training Materials, and Video

When the BSID-II is administered by trained clinicians and researchers, it can appear chaotic because the assessor moves flexibly through the items, clustering items with similar materials together, taking advantage of the child's waxing and waning attention to present items of interest, or re-presenting items to which the child was not attending on first presentation. Although the BSID-II items are numbered in the order of their increasing difficulty, the order of item presentation is not fixed. In order to maintain this degree of flexibility, the assessor must have the administration and scoring of every item memorized. This was not feasible for implementation in the field by ECLS-B interviewers. Therefore, the BSID-II administration booklet and scoring sheets were restructured for the 9-month data collection with the production of the Child Activity Booklet. For further information about this restructuring for the 9-month data collection, please refer to the *ECLS-B Methodology Report for the Nine-Month Data Collection, Volume 1: Psychometric Characteristics* (NCES 2005–100) (Andreassen and Fletcher 2005). The formatting used at 9 months was adopted for the 18-month field test in order to simplify the administration and to increase the clarity of the scoring criteria.

BSF-R administration was simplified by folding both administration instructions and score sheets into a single booklet and standardizing the formatting of each item to maximize efficiency. The item administration instructions and scoring criteria as presented in the BSID-II manual were closely adhered to while streamlining and making the administration and scoring as explicit as possible for field interviewers.

The application of the basal and ceiling rules was also simplified for the BSF-R. In the full BSID-II, items are arranged in age sets. The tester is instructed to administer additional age item sets depending on the numbers of credits and no credits the child receives and to continue administering additional age sets until the criterion has been satisfied. For example, on the BSID-II mental scale, if the child receives credit for fewer than five items (in the child's age set) then basal items (the next younger age set) should be administered. Conversely, if the child received no credit for three or more items, then ceiling items (the next older age set) should be administered.

For the BSF-R, the basal and ceiling rules were simplified so that interviewers only had to add up the number of credits, rather than having to keep track of both credits and no credits. On the mental scale, if the child received only 0 to 4 credits, then the interviewer would administer the set of basal items; if the child received 14 or more credits, then the interviewer would administer the set of

ceiling items. On the motor scale, if the child received only 0 to 4 credits, then the interviewer would administer the set of basal items, and if the child received 12 or more credits, then the interviewer would administer the ceiling items.

To improve the clarity of the administration and scoring instructions, a more structured layout was created for each item: item number and name across the top, with a picture of the materials used immediately underneath (exhibit 2-4). Below this information is a header labeled “Administration,” which includes the number of permissible administrations when more than one administration is allowed, and administration instructions just beneath.

The administration instructions were made as explicit as possible, with additional steps inserted to remind interviewers to look for a specific response or behavior at a specific time. For example, the last instruction on this sample page is to listen to and record whatever the child says as the interviewer reads the book.

Where appropriate, boxes were also included that gave explicit warnings, such as “Don’t let the child put the beads in his mouth,” as well as troubleshooting instructions for problematic situations that can arise.

The scoring criteria were highlighted in the box at the bottom, and special instructions were included to cover any special situations. For example, if the child were to build a tower of 8 blocks on the first try, he or she would automatically be given credit for building a tower of 6 blocks.

The score sheets, one for the mental scale and one for the motor scale, were on pullout sheets that could be folded over the administration pages so that the instructions were visible and the score boxes were handy. This improved upon the original BSID-II design in which the score sheets are entirely separate from the administration instructions. In addition, in the original BSID-II, the recommended order of item administration is different from the order in which items are listed on the score sheet. In the ECLS-B, a consistent numbering system was used in which items were administered in the same order in which they were listed on the score sheet.

1. Attends to Story



Administration

Listen for child's speech while attending to the book.

1. Place book on the table in front of the child.
2. Open it to the first page and say:
Look! See!
3. Let child explore the book, look at the pictures and turn the pages.
4. Then say:
Let's read the story.
5. Reposition yourself so you're sitting next to child.
6. Take book from child, open it, and begin to read, say:
Listen to the story.
7. Listen for child's response while you read and record above what child says.

| | |
|---|---|
| DID CHILD TALK IN RESPONSE TO BOOK? <input type="checkbox"/> Check box if child said AT LEAST two 2-word sentences. | Record what child said here: 1. _____ 2. _____ |
|---|---|

Scoring – Give credit if child...

1. Attends to entire story.

Attending includes decreasing motor activity and looking at the pictures, listening to the words, or talking to you about the pictures as you read.

The Bayley Short Form – Research Edition was developed for the Early Childhood Longitudinal Study, Birth Cohort and was adapted from the Bayley Scales of Infant Development-Second Edition. ©1993 by The Psychological Corporation, a Harcourt Assessment Company. Adapted and reproduced by permission. All rights reserved.

Bayley Scales of Infant Development is a registered trademark of The Psychological Corporation.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month data collection, 2003–04.

2.1.9 Identification of BSF-R Problem Items in the 18-Month Field Test

After the BSF-R data were collected in the 18-month field test, IRT analysis was used to determine whether the items performed as expected. The 18-month BSF-R was completed in 98 percent of the cases in the field test, offering an ample dataset for the IRT analyses. However, the age distribution of children in the 18-month field test was somewhat skewed. Most of the children in the 18-month field test were younger than the target age of 18 months.¹⁰ The age distribution is summarized in table 2-2. Because of the skewed age distribution, very few children in the 18-month field test were administered the ceiling items for either the mental scale (only 6 percent) or the motor scale (only 1 percent). Therefore, it was not possible to evaluate how well these ceiling items performed in the field test. The basal and core set items, however, were well represented and were thoroughly evaluated. Approximately 35 percent of children in the field test were administered the mental basal items and 23 percent were administered the motor basal items. This over-representation of children receiving the basal items sets is undoubtedly due to the skewed age range.

Table 2-2. 18-month field test age distribution of children: 2001

| Age in months | Percent |
|---------------|---------|
| 12–14 | 1.3 |
| 15 | 13.4 |
| 16 | 38.6 |
| 17 | 26.3 |
| 18 | 11.8 |
| 19 | 5.8 |
| 20 | 2.2 |
| 21–22 | 0.6 |

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 18-month field test; Spring 2001.

¹⁰ The sample for the field test was chosen in the same way as the main study sample; it included a sample of births that occurred in January through April 2000 within 15 PSUs in 7 states. The young age of the 18-month field test sample is a result of changes to and scheduling of the field tests for multiple rounds of collection. Some major issues related to interview length and ease of assessment administration were encountered during the field test for the first wave of data collection, conducted when the ECLS-B children were about 9-months-old. Significant changes were made to the 9-month instruments in order to address these issues, and a second field test for the 9-month data collection, which was not originally planned, was conducted to test the changes. Rather than select a new sample for the 18-month field test, cases that were originally sampled for the second 9-month field test were used for efficiency and because of time constraints, given that the collections were so close together. The 18-month field test began in June 2001 and went through November 2001. Thus, the sampled children were between about 13 and 18 months of age when the field test began.

Therefore, prior to conducting item calibration for the 18-month BSF-R items, it was necessary to weight the observations in the 18-month field test so that the age distribution would resemble the 18-month BSID-II standardization sample. Item calibration and scale equating established that a number of items did not perform as expected, as described in the following paragraphs.

On the mental scale, field test data fit the IRT 2-PL model quite well for all of the core items, with the exception of Men123, Builds tower of 6 cubes, and all of the basal items. Five items in the mental scale ceiling item set were identified as having poor calibration against the publisher data. Of these five, four had insufficient sample size to determine if they fit the IRT model. The mental ceiling items with poor fit include the following:

- Men129, Makes a contingent utterance;
- Men135, Builds tower of eight cubes;
- Men137, Matches four colors;
- Men141, Understands concept of one; and
- Men142, Produces multiple-word utterances in response to picture book.

Out of the total of 36 mental items, the above five items (14 percent) were excluded from the mental scale due to item-to-scale correlations below 0.20. However, as mentioned, the mental ceiling item set was administered to a small number of cases due to the skewed age range, 38 to 56 observations, depending on the item. This made it difficult to determine if the data for those four items really fit the IRT model. Only Men142 had a sufficient sample size ($n = 585$) to determine if it fit the model. However, the age range for this item specified in the manual is from 20 to 28 months of age. As a result, the probability value of this item in the ECLS-B was less than 3 percent, meaning that fewer than 3 percent of the 585 scores were credits.

On the motor scale, field test data fit the IRT model well for all of the basal set items. Three items in the core set did not calibrate well against publisher data and had item-to-scale correlations below 0.20. These items were not included in the calibration but were included in the scale. These three core items include the following:

- Mot074, Uses pads of fingertips to grasp pencil;
- Mot087, Jumps distance of 4 inches; and

Mot089, Walks on tiptoe for four steps.

Due to the skewed age distribution, the nine motor ceiling items were administered to only seven or eight toddlers, depending on the item. Two of the ceiling items could not be calibrated because none of the children received credit, however, they were retained in the scale. These ceiling items on which no one received credit include Mot100, Stops from a full run, and Mot101, Buttons one button.

Three additional items in the motor ceiling set were excluded from the scale due to item-to-scale correlations below 0.20. These items included the following:

Mot093, Manipulates pencil in hand;

Mot099, Walks on tiptoe for 9 feet; and

Mot103, Stands alone on right foot for 4 seconds.

The remaining four ceiling items showed acceptable fit despite the small sample size.

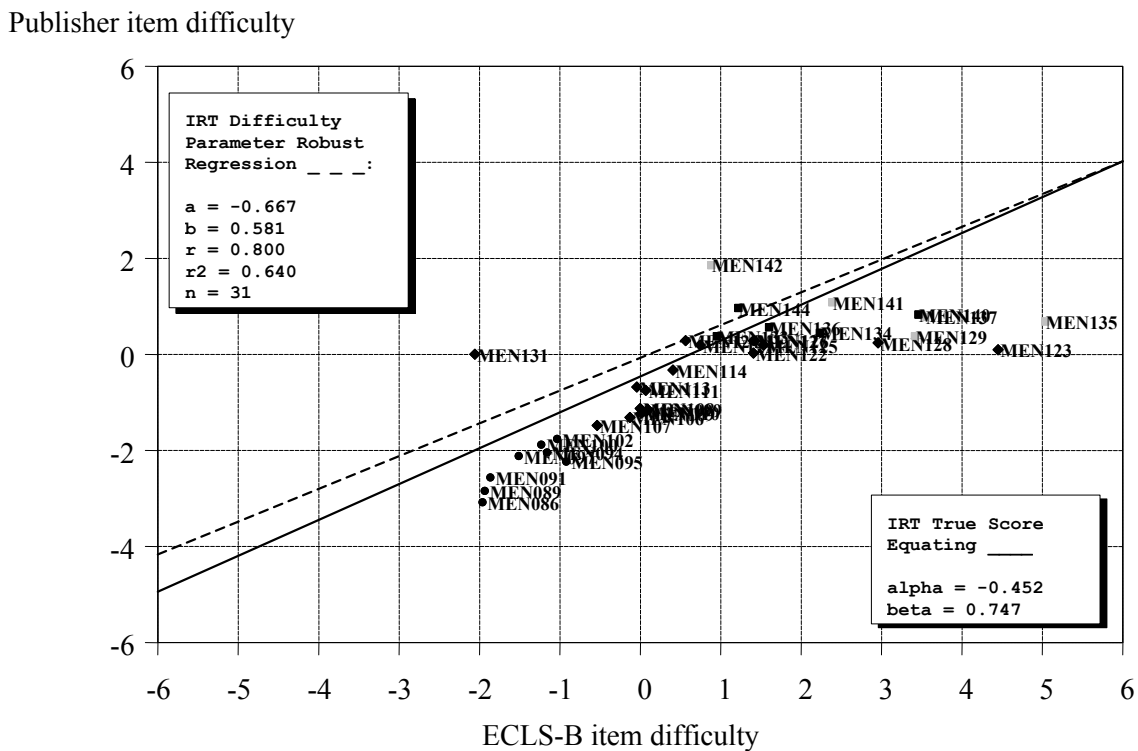
Of these excluded items, “Uses pads of fingertips to grasp pencil” was of the greatest concern. As a fine motor item, it should have been less sensitive to the skewed age distribution. It would appear from the data that interviewers were uncertain how to score this item and suggested that training on this item had to be clarified.

The remaining items on the motor scale that were identified as problematic were also sensitive to the skewed age distribution and the emerging motor skills of children. At 18 months, toddlers are just beginning to jump off the floor, walk on tiptoe, and stand on one foot. The children in the field test were probably too young for these items, therefore, the results obtained are not a fair indicator of what children at the correct target age can do. Nevertheless, IRT analyses of publisher data were conducted in order to identify items at roughly equal intervals of difficulty with discrimination parameters that show the items successfully differentiate those who can perform the activities identified in the items from those who cannot. The goal of these analyses was to assemble the best possible items for the BSF-R. That having been accomplished, the goal of piloting the 18-month BSF-R in the field test was to identify misbehaving items by conducting IRT analyses of field test data. Items that perform well would cluster linearly around the regression line of publisher data and field test data. Items that do not perform well would be more distant from the regression line. The following two figures demonstrate how well the

BSF-R mental and motor items from the field test cluster around their respective regression lines with publisher data.

Figure 2-20 demonstrates that the 18-month BSF-R mental scale items worked rather well, with the exception of the earlier-mentioned ceiling items. The location of Men131 (attends to story) suggested that interviewers too often gave credit for it; however, since this could be corrected by focusing more attention on the scoring criteria during training, this item was not considered problematic.

Figure 2-20. Item Response Theory (IRT) equating of 18-month BSF-R mental scale field test data and publisher standardization data: 1993



NOTE: a = regression intercept, when publisher item difficulty parameters are regressed on BSF-R item difficulty parameters; b = slope coefficient, when publisher item difficulty parameters are regressed on BSF-R item difficulty parameters; alpha = linear transformation of scale; beta = linear transformation of origin. Both lines represent a linear regression of the ECLS-B source test on the publisher target test. The simple linear regression $y = a + bx$ yields intercept coefficient a and a slope coefficient b. The IRT difficulty parameter robust regression (dashed line) represents the best linear transformation of the ECLS-B data and slope required to place ECLS-B IRT difficulty parameters on the same scale metric as those of the publisher. IRT true-score equating (solid line) represents the best linear transformation of the ECLS-B and slope required to match the ECLS-B test characteristic curve as closely as possible with that of the publisher.

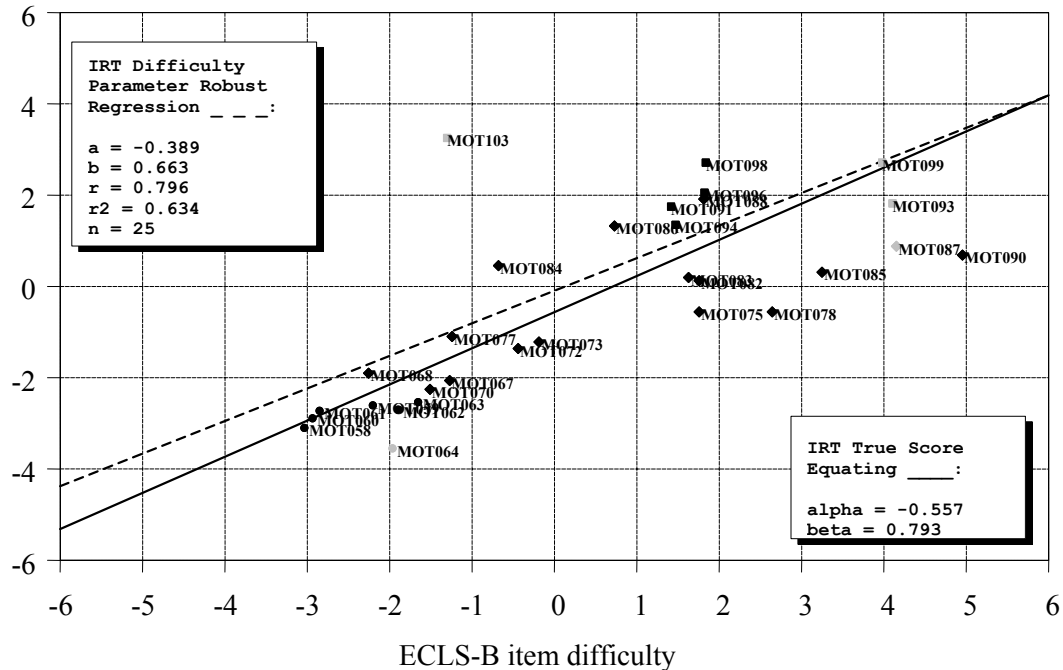
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 18-month field test (2001) and Psychological Corporation publisher dataset for the Bayley Scales of Infant Development, Second Edition, 1993.

Figure 2-21 shows that the BSF-R motor scale also had some problem items, as evidenced by the wider scatter of items throughout. However, it should be kept in mind that the motor items are

probably more sensitive to physical development. The children in the field test were probably too young for these items. Therefore, it is not possible to determine conclusively that the items were problematic.

Figure 2-21. Item Response Theory (IRT) equating of 18-month BSF-R motor scale field test data and publisher standardization data: 1993

Publisher item difficulty



NOTE: a = regression intercept, when publisher item difficulty parameters are regressed on BSF-R item difficulty parameters; b = slope coefficient, when publisher item difficulty parameters are regressed on BSF-R item difficulty parameters; alpha = linear transformation of scale; beta = linear transformation of origin. Both lines represent a linear regression of the ECLS-B source test on the publisher target test. The simple linear regression $y = a + bx$ yields intercept coefficient a and a slope coefficient b. The IRT difficulty parameter robust regression (dashed line) represents the best linear transformation of the ECLS-B data and slope required to place ECLS-B IRT difficulty parameters on the same scale metric as those of the publisher. IRT true-score equating (solid line) represents the best linear transformation of the ECLS-B and slope required to match the ECLS-B test characteristic curve as closely as possible with that of the publisher.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 18-month field test (2001) and Psychological Corporation publisher dataset for the Bayley Scales of Infant Development, Second Edition, 1993.

More detailed IRT analysis of the 18-month field test was not conducted due to the discontinuation of the 18-month national data collection. Instead, the decision was made to combine the 18- and 30-month data collections into a single 2-year data collection, and efforts were redirected to the development of the 2-year BSF-R.