3. DEVELOPING THE 2-YEAR BAYLEY SHORT FORM–RESEARCH EDITION

After the decision was made to combine the 18- and 30-month data collections into a single data collection at 2 years, additional work was done to extend the 18-month Bayley Short Form–Research Edition (BSF-R) upward to form the basal items and lower core items for a 2-year BSF-R, and to incorporate items from the 30-month BSF-R to form the upper end of the core items and the ceiling items. Chapter 2 summarized the extensive Item Response Theory (IRT) analysis that was done to identify the pool of candidate items for the 18-month BSF-R and the feasibility criteria that guided the further selection of items that could be used in a field setting. This chapter describes how the work that was done to develop the 18-month version of the BSF-R was used as the foundation to develop the 2-year BSF-R, while maintaining the administration and scoring standards of the Bayley Scales of Infant Development, Second Edition (BSID-II). (Please see section 2.1.1 for a description of the BSID-II.)

3.1 Pilot Testing of 2-Year BSF-R and Results

With the shift to a data collection at 2 years instead of at 18 months, the results of the IRT analysis described in chapter 2 offered a theoretical starting point for revisions to the BSF-R to make it appropriate for the new target age. Some of the basal items and the lower half of the core items from the 18-month BSF-R were useful as basal items at 2 years. Some of the upper half of the 18-month core items and the ceiling items were suitable for the lower half to middle of the core items. Items from the basal and core sets of the 30-month version were suitable for the middle to upper end of the 2-year core set and for the ceiling set.

Although the items for the 30-month BSF-R had been identified already and work had begun to reformat the administration and scoring instructions to make them consistent with the Child Activity Booklet (CAB) used at 9 months and 18 months, the 30-month version had not been field tested yet. This meant that items at the upper ranges of ability, corresponding to the upper half of the core and the ceiling items, first had to be confirmed as appropriate for use at 2 years and had to be tested for operational feasibility in a field setting.

The first step was to confirm that the candidate 30-month items were appropriate for the 2-year core and ceiling sets on both the mental scale and the motor scale and that candidate items in the 18-

month BSF-R were appropriate for the basal set and lower core items. Items that had been field tested, or at least pilot-tested, and had proven their feasibility in the field were more likely to be included in the 2-year BSF-R than those items that had not.

The publisher-recommended 23- to 25-month age set was the starting point for both the mental scale and the motor scale. The ability parameters for the items in the BSID-II mental scale 23- to 25-month age set ranged from 2.708 to 7.020, and the ability parameters for the items in the 30-month basal items began at 2.708. Therefore, the 30-month BSF-R mental basal items were at the appropriate level of difficulty for 2-year core items. (In fact, during the design of the 30-month BSF-R, the 23- to 25-month age set had been identified by IRT analysis as the best source for basal items at 30 months.) Please see section 2.1.3 for an explanation of how to interpret ability parameters.

Because the ability parameters for the items in the BSID-II motor scale 23- to 25-month age set ranged from 1.677 to 3.038, this age set was identified as the best source for BSF-R motor scale basal items at 30 months because they are beyond 1 standard deviation below the mean of the ability level for that age set. Therefore, the motor basal items in the 30-month BSF-R, which ranged from 2.102 to 3.249, were suitable for upper core items at 2 years. Additionally, BSID-II items with ability parameters that were within 1 standard deviation above and 1 standard deviation below the mean of the ability parameter for this age item set were also identified and targeted for the basal and ceiling items.

Thereafter, the ability parameters for items 2 to 3 standard deviations above and below the mean for the publisher defined 23- to 25-month age set were identified. These items were targeted for the outermost tails of the basal item set and the ceiling item set.

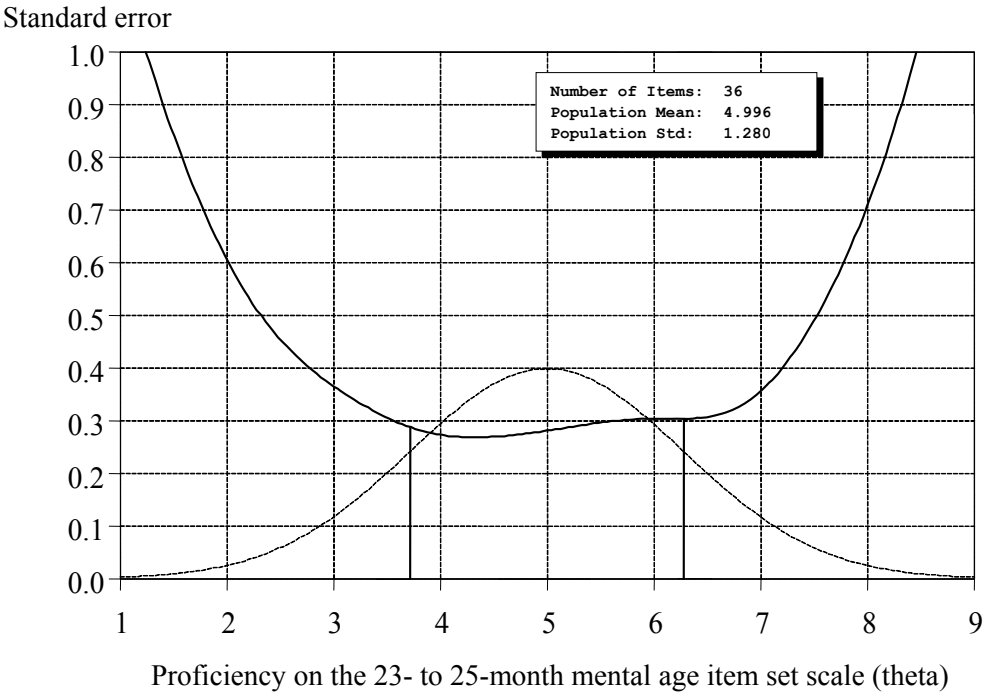
Once a list of candidate items that ranged from -3 to +3 standard deviations around the mean of the ability distribution for 2-year-olds was composed, items that were clearly not operationally feasible were eliminated. On the mental scale, for example, the item “Identifies objects in photographs” was eliminated because too many materials were required and the administration was complicated. On the motor scale, all stair items (e.g., “Walks upstairs with help”) were eliminated because they required a double set of three steps (i.e., three steps up, platform, then three steps down) constructed to the standard overall dimensions of 60” (depth) x 24” (width) x 19 ½” (height). Interviewers would have to bring these stairs on home visits, which was not feasible in the field.

The second step was to conduct a small pilot study to test these items in a home visit setting that replicated, to the extent possible, the context of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) home visit. Working in teams of two, Westat child development staff members assessed children with varying combinations of candidate BSF-R items. After each home visit, these staff members recorded their impressions and discussed them during weekly debriefing meetings. The administration and scoring of items and their feasibility for field staff were also discussed during those meetings. In this way, difficult items were eliminated and replaced with likely candidates (with good psychometrics) on a rolling basis. Of the 36 items in the BSID-II mental scale 23- to 25-month age set, 17 were retained for the core item set; and of the 19 items in the BSID-II motor scale 23- to 25-month age set, 17 were retained for the core item set.

3.2 Psychometrics of the 2-Year BSF-R Mental Scale Core Item Set

The 2-year BSF-R mental scale was constructing using the same IRT 2-parameter logistic (2-PL) analysis and item selection criteria as in previous versions of the BSF-R. For comparison purposes, figure 3-1 demonstrates the standard error of the BSID-II mental scale 23- to 25-month age set, which shows a standard error below or at 0.3 for children scoring between 1 standard deviation below and 1 standard deviation above the mean based on the standardization dataset. This is the target standard error that the BSF-R mental core item set should ideally achieve.

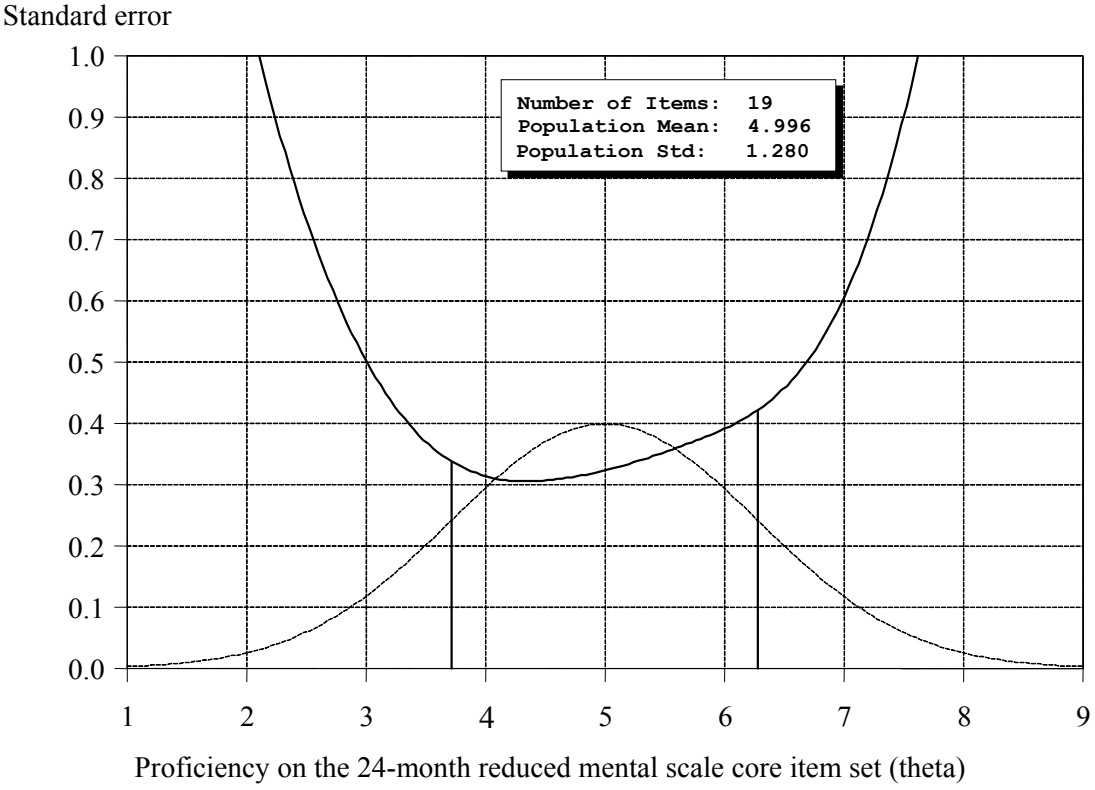
Figure 3-1. Standard error of measurement by proficiency level for the publisher-recommended 23- to 25-month age item set of the mental scale: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



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

Figure 3-2 demonstrates the standard error that would be expected from the BSF-R mental scale core item set that was created. This is an estimate based on the publisher standardization dataset. This graph shows that the expected standard error would be less than 0.4 for most of the core, only slightly exceeding 0.4 at approximately 1 standard deviation above the mean.

Figure 3-2. Standard error of measurement by proficiency level for the 2-year BSF-R mental scale core item set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



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

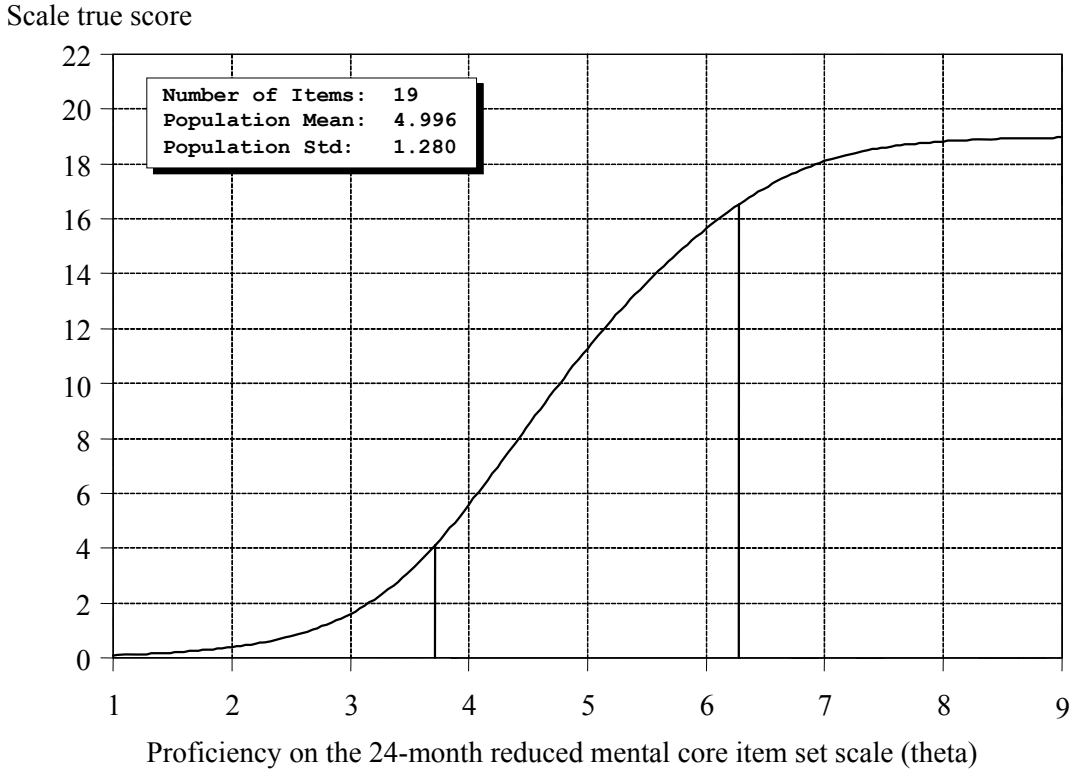
3.3 Creating the 2-year BSF-R Mental Scale Basal and Ceiling Sets

Attention was then turned to the final selection of items for the basal item set and the ceiling item set. Items with ability parameters from 1 to 3 standard deviations below the mean were reviewed and those with the best discrimination parameters were selected as candidate items for the basal item set. To select items for the ceiling item set, items from 1 to 3 standard deviations above the mean were reviewed and those with the best discrimination parameters were selected as candidate items for the ceiling item set. Item feasibility for administration in the field by field staff was then evaluated.

To evaluate candidate basal items, children up to 6 months younger than 2 years were assessed. This younger age was selected in order to test the lower limits of the items. To evaluate candidate items for the upper range of the core items and for the ceiling item sets, children as much as 6 months older than the target age were assessed. These older ages were selected in order to test the upper limits of the items. Once the best items for the core set, basal set, and ceiling set were identified, optimal ordering of items was tested using varying combinations of items. In total, the mental and motor items from preliminary versions of the 2-year BSF-R and the ordering of the items were tested on approximately 40 to 45 children.

Once the candidate items for the mental scale were selected, it was then necessary to determine the basal and ceiling rules that would route children to those supplementary item sets, as necessary. The same procedure that was followed when designing the previous versions of the BSF-R (see chapter 2) was followed again and is demonstrated in figure 3-3, which shows IRT true scores by ability for the BSF-R mental core item set. The vertical line on the left shows that children scoring from 0 to 4 on the core items should be administered the basal set of items. The vertical line on the right shows that children scoring from 16 to 19 on the mental core items should be administered the ceiling set of items.

Figure 3-3. Establishing basal and ceiling rules for the 2-year BSF-R mental core item set using true scale scores: IRT 2-parameter logistic item calibrations using publisher data: 1993



NOTE: Vertical lines descending from the solid curve to the x-axis indicate 1 standard deviation above the mean. Std = standard deviation.
SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

3.4 Content of 2-Year BSF-R Mental Scale

On the basis of IRT 2-PL analysis and pilot testing to assess item feasibility in a field setting, the items listed in exhibit 3-1 were selected for the 2-year BSF-R mental scale. This exhibit also summarizes the materials needed for each item, the BSID-II item number and item description, and the age range for which the item is appropriate. The third column in this exhibit indicates whether an item had to be administered in order to obtain a score, or whether it could be scored either from the administration of another item or from observation of the child's spontaneous behavior.

Exhibit 3-1. 2-year BSF-R mental scale items, materials, item descriptions, and item age ranges: 2003–04

Item numbers	Item description	Material	Number of adminstrations[1]	Age range (months)[2]
		Total core set	14	
		Total basal set	2–3	
		Total ceiling set	6	
		Basal item set		
Men099	Points to two pictures	Stimulus page	(Core score)[3]	12–19
Men106	Uses words to make wants known	None needed	Observe	14–19
Men107	Follows directions	Doll	1	14–22
Men108	Points to three of doll's body parts	Doll	1	14–22
Men109	Names one picture	Stimulus page	(Core score)[3]	14–22
Men110	Names one object	Ball, cup, etc.	(Core score)[3]	14–22
Men111	Combines word and gesture	None needed	Observe	14–22
Men113	Says eight different words	None needed	Obs/adm[4]	17–25
Men114	Uses a two-word utterance	None needed	Observe	17–25
		Core item set		
Men117	Imitates a two-word sentence	None needed	Obs/adm[4]	17–25
Men121	Uses pronouns	None needed	Observe	17–25
Men122	Points to five pictures	Stimulus page	1	17–25
Men123	Builds tower of six cubes	Cubes	1	17–28
Men124	Discriminates book, cube and key	Book, cube, key	1	17–28
Men125	Matches pictures	Stimulus page	1	17–28
Men126	Names three objects	Book, block, etc.	1	17–28
Men127	Uses a three-word sentence	None needed	Observe	17–28
Men128	Matches three colors	Stimulus page	1	20–28
Men131	Attends to story	Book	1	20–31
Men133	Names five pictures	Stimulus page	Joint, 122	20–31
Men134	Displays verbal comprehension	Stimulus page	1	20–31
Men135	Builds tower of eight cubes	Cubes	Joint, 123	20–31
Men136	Poses questions	None needed	Observe	23–31
Men137	Matches four colors	Stimulus page	1	23–31
Men140	Understands two prepositions	Cups, bunny	1	23–34
Men141	Understands concept of one	Cubes	1	23–37
Men144	Discriminates pictures I	Stimulus page	1	23–37
Men145	Compares sizes	Stimulus page	1	23–37

See notes at end of exhibit.

Exhibit 3-1. 2-year BSF-R mental scale items, materials, item descriptions and item age ranges: 2003–04 —Continued

Item numbers	Item description	Material	Number of adminstrations[1]	Age range (months)[2]
		Ceiling item set		
Men142	Produces multiword utterances to book	Book	Observe	23–37
Men146	Counts (number names)	Cubes	1	23–42
Men147	Compares masses	Blue boxes	1	23–47
Men148	Uses past tense	None needed	Observe	23–42
Men151	Discriminates pictures II	Stimulus page	1	26–42
Men152	Repeats three number sequences	None needed	1	26–42
Men153	Understands four prepositions	Cups, bunny	(Core score)[3]	26–42
Men154	Identifies gender	None needed	1	26–42
Men162	Sorts pegs by color	Pegs, bags	1	32–42

[1] This column counts the number of administrations that are done. Each administration is defined as the structured presentation of the stimulus material to obtain the child's response. Thus, multiple items that are scored with the same administration only count as one total administration. These items are indicated by "joint" followed by the number of the paired item. The actual number of items may be less important for determining the overall time burden than the number of different administrations required.

[2] The age ranges are based on the youngest and oldest item sets in the original BSID-II. An age range of 14–22 months indicates the item is included in the 14-month through the 22-month age sets of the BSID-II.

[3] Core score means that the score for this item is obtained during the administration of a core item.

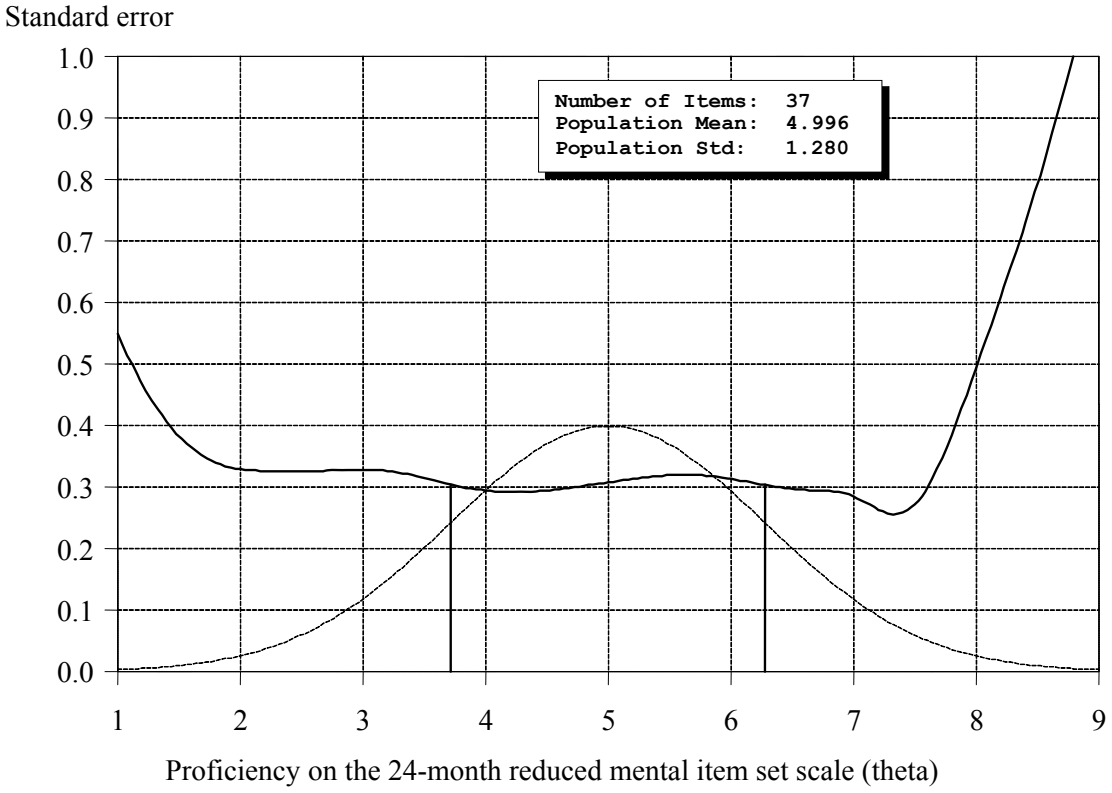
[4] Some items can be scored by observation during testing, but if the item is not observed, then it is administered.

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.

3.5 Projected Standard Error of Measurement for the 2-Year BSF-R Mental Scale

With the above selected items and implementation of the above decision rules, it was possible to estimate the forecast reliability of the 2-year BSF-R mental scale. To reach a forecast reliability of 0.80, which was the target recommended by the IRT experts on the assessment work group panel, it would be necessary to have an expected standard error of 0.4 or less. Figure 3-4, shows that for the mental scale, this target would be achieved, based on the publisher standardization dataset.

Figure 3-4. Projected standard error of measurement by proficiency level for the 2-year BSF-R mental scale: IRT 2-parameter logistic item calibrations using publisher data: 1993

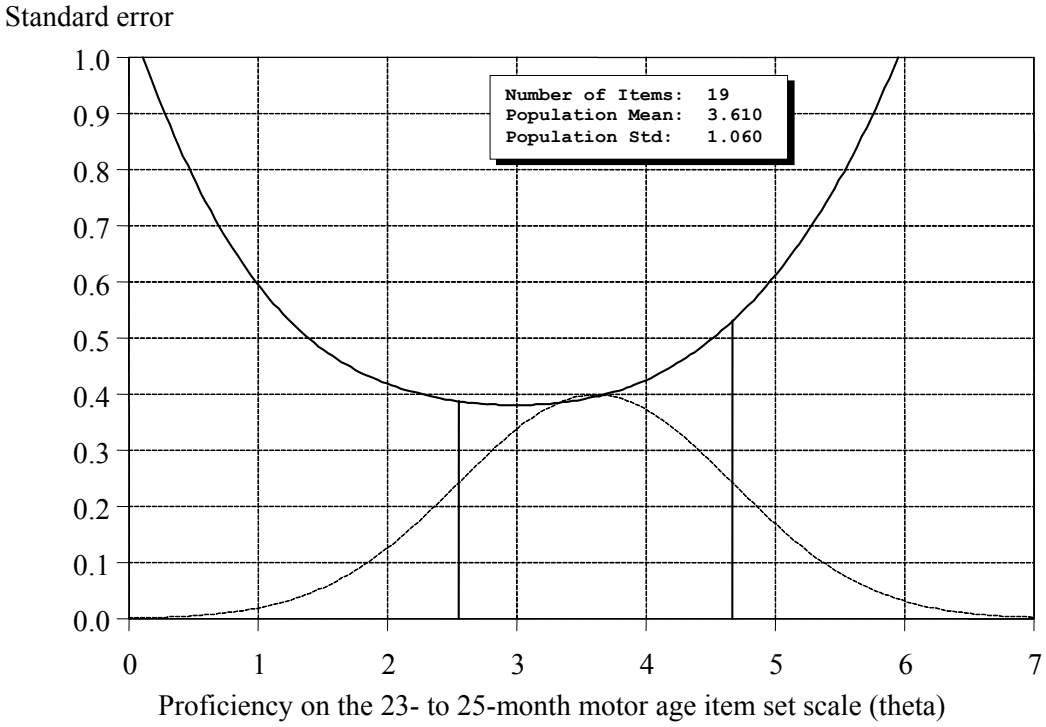


NOTE: Vertical lines descending from the solid curve to the x-axis indicate 1 standard deviation above the mean. Std = standard deviation. The normal curve (dashed line) represents the projected latent distribution of the 23- to 25-month population and is included for illustration purposes.
SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

3.6 Psychometrics of the 2-Year BSF-R Motor Scale Core Item Set

The 2-year BSF-R motor scale was constructed using the same procedures as for previous versions of the BSF-R and for the 2-year mental scale. For comparison purposes, figure 3-5 demonstrates the standard error of the BSID-II motor scale 23- to 25-month age set, which shows a standard error below or at 0.3 for children scoring between 1 standard deviation below and 1 standard deviation above the mean based on the standardization dataset.

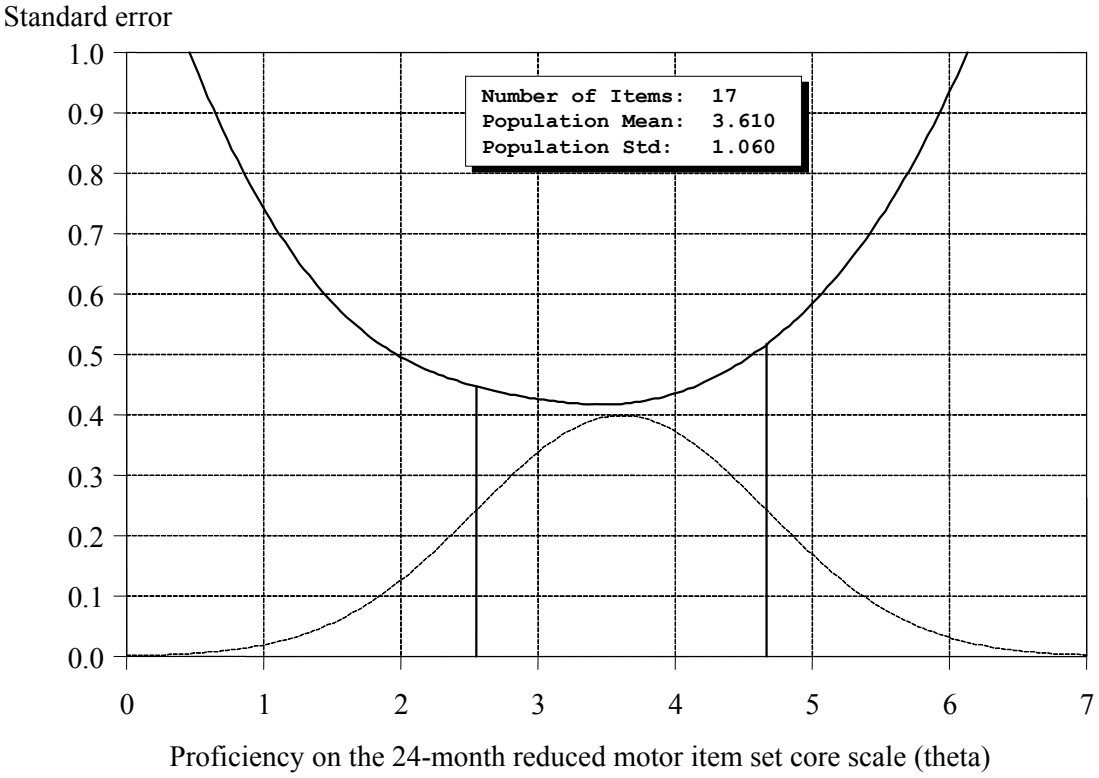
Figure 3-5. Standard error of measurement by proficiency level for the BSID-II motor scale 23- to 25-month age item set: IRT 2-parameter logistic item calibrations using publisher data: 1993



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

Figure 3-6 demonstrates the standard error that would be expected from the BSF-R motor scale core item set if administered under conditions similar to the BSID-II. This is an estimate based on the publisher standardization dataset. This graph shows that the expected standard error would be less than 0.4 for a good part of the core, exceeding 0.4 at approximately 1 standard deviation above the mean.

Figure 3-6. Projected standard error of measurement by proficiency level for the 2-year BSF-R motor scale core item set: IRT 2-parameter logistic item calibrations using publisher data: 1993



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

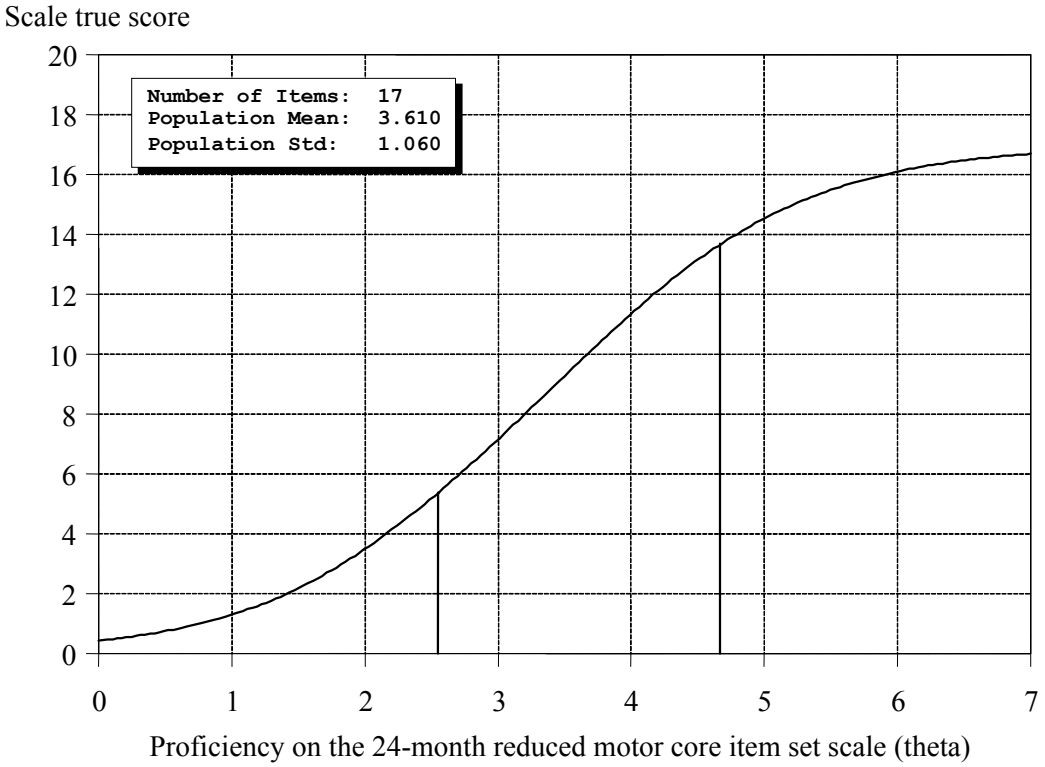
3.7 Creating the 2-Year BSF-R Motor Scale Basal and Ceiling Sets

Following the same procedures as for the mental scale basal and ceiling sets, the items for the basal set and for the ceiling set were selected. Items with ability parameters from 1 to 2 standard deviations below the mean were reviewed and those with the best discrimination parameters were selected as candidate items for the basal item set. To select items for the ceiling item set, items from 1 to 2 standard deviations above the mean were reviewed and those with the best discrimination parameters were selected as candidate items for the ceiling item set. Item feasibility for administration in the field by field staff was then evaluated.

To evaluate candidate basal items, children up to 6 months younger than 2 years were assessed. This younger age was selected in order to test the lower limits of the items. To evaluate candidate items for the upper range of the core items and for the ceiling item sets, children as much as 6 months older than the target age were assessed. These older ages were selected in order to test the upper limits of the items. Once the best items for the core set, basal set, and ceiling set were identified, optimal ordering of items was tested using varying combinations of items.

Once all the items for the motor scale were selected, it was then necessary to determine the basal and ceiling rules that would route children to those supplementary item sets, as necessary. The same procedure that was followed when designing the previous versions of the BSF-R was followed again and is demonstrated in figure 3-7, which shows IRT true scores by ability for the BSF-R motor core item set. The vertical lines in figure 3-7 indicate 1 standard deviation above and below the mean. Children with scores beyond 1 standard deviation were administered either the basal or ceiling items, as appropriate. The vertical line on the left shows that children scoring from 0 to 4 on the motor core items should be administered the basal set of items. The vertical line on the right shows that children scoring from 13 to 17 on the motor core items should be administered the ceiling set of items.

Figure 3-7. Establishing basal and ceiling rules for the 2-year BSF-R motor core item set using true scale scores: IRT 2-parameter logistic item calibrations using publisher data: 1993



NOTE: Vertical lines descending from the solid curve to the x-axis indicate 1 standard deviation above the mean. Std = standard deviation.
SOURCE: Publisher standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

3.8 Content of the 2-Year BSF-R Motor Scale

On the basis of IRT 2-parameter logistic analysis and pilot testing to assess item feasibility in a field setting, the items listed in exhibit 3-2 were selected for the 2-year BSF-R motor scale. This exhibit also summarizes the materials needed for each item, the BSID-II item number and item description, and the age range for which the item is appropriate. The third column in this exhibit indicates whether an item had to be administered in order to obtain a score, or whether it could be scored either from the administration of another item or from observation of the child's spontaneous behavior.

Exhibit 3-2. 2-year BSF-R motor scale items, materials, item age ranges, and item descriptions: 2003–04

Item numbers	Item description	Material	Number of administrations[1]	Age range (months)[2]
		Total core set	14	
		Total basal set	3 to 6	
		Total ceiling set	3	
		Basal item set		
Mot059	Stands up I	None needed	(Core score)[3]	8–12
Mot062	Walks alone	None needed	Obs/adm[4]	9–13
Mot063	Walks alone with good coordination	None needed	Obs/adm[4]	11–16
Mot065	Squats briefly	Red ball	Obs/adm[4]	11–16
Mot068	Stands up II	None needed	(Core score)[3]	11–19
Mot070	Grasps pencil at middle	Pencil, paper	1	12–22
Mot072	Stands on right foot with help	Squeaky toy	1	12–22
Mot073	Stands on left foot with help	Squeaky toy	Joint 072	13–22
Mot077	Runs with coordination	Red ball	1	14–25
		Core item set		
Mot075	Uses hand to hold paper in place	Pencil, paper	1	13–25
Mot074	Uses pads of fingertips to grasp pencil	Pencil, paper	Joint 075	13–22
Mot078	Jumps off floor (both feet)	Tape measure	1	14–25
Mot082	Stands alone on right foot	Squeaky toy	1	17–28
Mot083	Stands alone on left foot	Squeaky toy	Joint 083	20–28
Mot084	Walks forward on line	Tape measure	1	20–31
Mot085	Walks backward close to line	Tape measure	1	20–31
Mot086	Swings leg to kick ball	Ball	1	20–31
Mot087	Jumps distance of 4 inches	Tape measure	1	23–31
Mot088	Laces three beads	Laces, 3 beads	1	23–34
Mot089	Walks on tiptoe for four steps	Tape measure	1	23–34
Mot090	Grasps pencil at nearest end	Pencil, paper	1	23–34
Mot091	Imitates hand movements	None needed	1	23–37
Mot092	Tactilely discriminates shapes	Bag, shapes	1	23–37
Mot093	Manipulates pencil in hand	Paper, pencil	Observe	23–37
Mot094	Stands up III	None needed	1	26–37
Mot096	Copies circle	Paper, pencil	1	26–42

See notes at end of exhibit.

Exhibit 3-2. 2-year BSF-R motor scale items, materials, item age ranges, and item descriptions: 2003–04 —Continued

Item numbers	Item description	Material	Number of administrations[1]	Age range (months)[2]
		Ceiling item set		
Mot098	Imitates postures	None needed	1	29–42
Mot099	Walks on tiptoe for 9 feet	Tape measure	(Core score)[3]	29–42
Mot101	Buttons one button	Button sleeve	1	29–42
Mot102	Stands alone on left foot for 4 seconds	Squeaky toy	(Core score)[3]	32–42
Mot103	Stands alone on right foot for 4 seconds	Squeaky toy	(Core score)[3]	32–42
Mot107	Hops twice on 2 feet	Tape measure	1	32–42

[1]This column counts the number of administrations that are done. Each administration is defined as the structured presentation of the stimulus material to obtain the child's response. Thus, multiple items that are scored with the same administration only count as one total administration. These items are indicated by "joint" followed by the number of the paired item. The actual number of items may be less important for determining the overall time burden than the number of different administrations required.

[2]The age ranges are based on the youngest and oldest item sets in the original BSID-II. An age range of 14–22 months indicates the item is included in the 14-month through the 22-month age sets of the BSID-II.

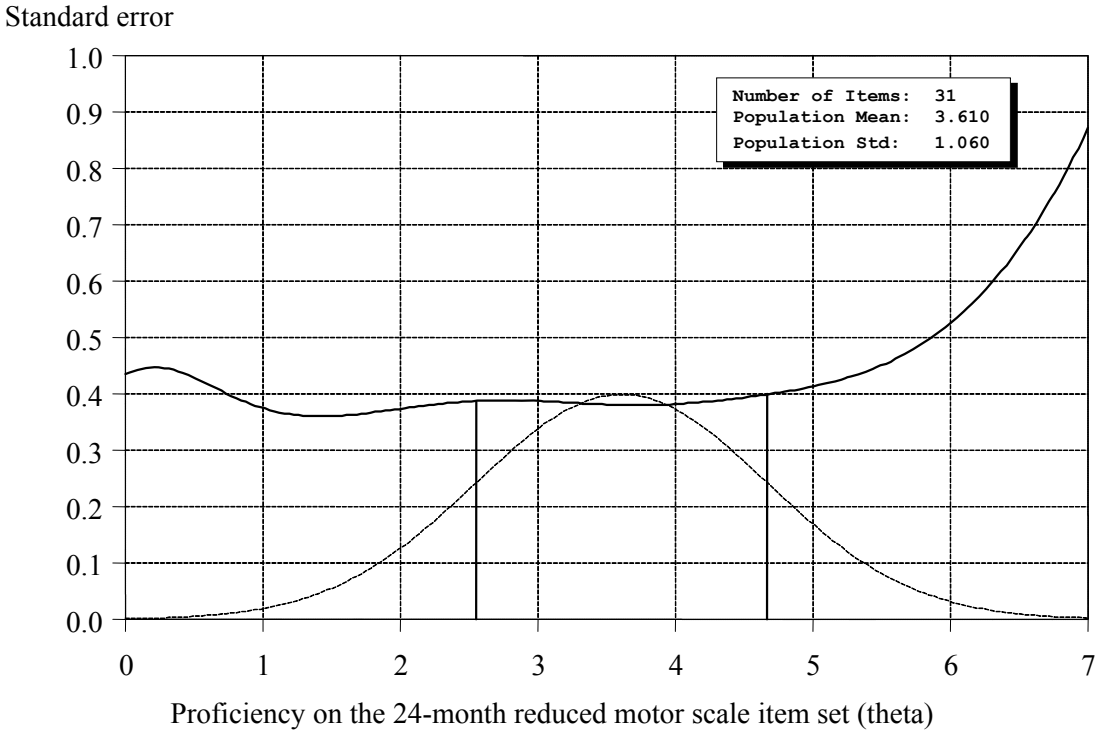
[3] Core score means that the score for this item is obtained during the administration of a core item.

[4] Some items can be scored by observation, but if the item is not observed, then it is administered.

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.

Based on the above selected items and implementation of the above basal and ceiling decision rules, it was possible to estimate the forecast reliability of the 2-year BSF-R motor scale. In order to reach a forecast reliability of 0.80, which was the target recommended by the IRT experts on the assessment workgroup panel, it would be necessary to have an expected standard error of 0.4 or less. Figure 3-8 shows that for the motor scale this target would be achieved for the most part, based on the publisher standardization dataset. However, beyond 3 standard deviations above the mean, this level of reliability would not be achieved. Although the BSID-II items were carefully reviewed, it was not possible either to replace any items with other items having higher discrimination parameters (and, therefore, bring down the curve) or to add any items to remedy this situation. This is partly due to not including any stair items and partly because the BSID-II does not have adequate item coverage at the upper end of the ability distribution, as evidenced by the paucity of items and the wide gaps in the ability parameters at the high end of the motor scale.

Figure 3-8. Projected standard error of measurement by proficiency level for the 2-year BSF-R motor scale: IRT 2-parameter logistic item calibrations using publisher data: 1993



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

3.9 Subsequent Revisions in Preparation for 2-Year National Training

Based on the review conducted of the results of the 18-month field test BSF-R and the small pilot study of the 2-year BSF-R, revisions were made to the administration instructions in the CAB. The only major revision involved replacing the shield in MOT092, "Tactilely discriminates shapes," with a small cloth bag. The shield is a thin sheet of opaque plastic about the size of a sheet of letter paper. At the midpoint on one of the long edges is a semicircular cutout. The shield is supposed to be placed over the child's wrist at the semicircle to obstruct the child's view of the item materials. Some children at this age find this frightening. To make this procedure easier for interviewers and to ease the apprehension of some children, the shield was replaced by a cloth bag. Before making this change, however, Dr. Kathleen Matula, an expert on the BSID-II who had worked on the restandardization of the BSID-II, was consulted. She confirmed that the shield is a problem and that substituting the cloth bag was a benign change that

should have no impact on children's performance on this item. Small improvements were also made to the administration and scoring instructions in the CAB. For example, the instructions specified that the administrator had to hold the button sleeve during the administration of the item "Buttons one button." However, pilot testing showed that children this age typically wanted to hold the button sleeve by themselves. Allowing that, however, increases the difficulty of the item because children this age do not have sufficient fine motor control both to hold the button sleeve and button the button. Therefore, it was necessary to re-emphasize the need for the administrator to hold the button sleeve in order to adhere to the standardized administration instructions in the BSID-II.

3.10 Training Procedures and Standardized Training Video

To ensure that all potential trainers for the 2-year national training had the same knowledge about the administration and scoring of the BSF-R, a standardized training videotape was produced that detailed the administration and scoring instructions for all items in the 2-year BSF-R. A similar standardized training videotape, produced for the 18-month field test training, had been well-received by trainees. However, at that training the trainees were not permitted to keep a copy of the training video. Instead, at the completion of training, an informal tutorial videotape that summarized the basics of item administration and scoring was sent to trainees. This videotape was also well received and interviewers reported that they found the tutorial helpful in mastering item administration and scoring. This led to the decision for the 2-year training to produce enough copies of the standard training videotape to enable all trainees to take them home and review them periodically as the need arose.

Dr. Kathleen Matula, who had been involved in the development and restandardization of the BSID-II while at The Psychological Corporation, reviewed the videotape and approved the content to verify that the information on the video was accurate. It was then shown to all interviewers during the training so that all field staff received the same information during the 2-year national training, which required approximately 10 rooms and, therefore, 10 lead trainers.

Dr. Matula also reviewed the 2-year BSF-R administrations of the four core child development staff to verify that their administrations were up to professional standards, their scoring was accurate, and their ability to build rapport with the children was strong. Thirteen additional staff members (designated as lead and assistant trainers) were then trained on the 2-year BSF-R and followed the same certification procedure as the one used to certify field staff at the national training, as described in the next section. The lead and assistant trainers were videotaped while administering the 2-year BSF-R to

children recruited from the community. The four core staff members then reviewed the videotapes using the same quality review form that would be used during the national training for field staff. All trainers passed this certification process with scores for administration for both the mental and motor scales averaging 100 percent. Average scores for scoring accuracy were 99 percent for the mental scale and 94 percent for the motor scale. Only core staff, lead trainers, and assistant trainers who were trained and certified on the BSF-R were permitted to review and score the videotapes of the trainees at the national training.

3.11 Certification Procedures for the 2-Year BSF-R

Since the BSF-R is a derivative of the BSID-II, a comprehensive, standardized measurement, administrators must follow the administration instructions clearly and adhere to strict criteria when scoring the child's performance on an item. The administrator also must establish and maintain good rapport with the child in order to elicit the child's best performance. Maintaining good rapport requires the interviewer to adjust his or her pace to match the child's ability to receive the instructions and to monitor the child's positive or negative mood to keep the child motivated. In order to monitor the child's performance and mood in this way, interviewers must be in command of the item administration procedures and scoring criteria.

For these reasons, it was important to ensure that, before administering the BSF-R in the field, trainees be able to administer the BSF-R to publisher standards for administration and scoring. Standards were developed by Westat's child development experts, with guidance from external reviewers. A three-level certification component was incorporated into training, beginning with in-class exercises, progressing to a written precertification exam, and culminating in the complete administration of the BSF-R during a "live" practice session with children. Certification on the BSF-R during the live practice session involved evaluating the trainees' ability to administer the items according to the standardized instructions, to apply the scoring criteria for each item, and to establish rapport and interpret children's responses.

Trainees took several in-class written quizzes during 2 days of direct instruction on the BSF-R. Beginning with the introduction of the mental items of the BSF-R on the BSF-R training videotape, trainees were quizzed on the scoring of all items, including core, basal, and ceiling items. The quizzes were collected and immediately reviewed during a break by trainers and assistant trainers. The purpose of this quiz was to identify individuals who were having difficulty understanding how to score

the items. Once individuals having difficulties were identified, one-on-one feedback and remediation were provided as needed before proceeding to the next level. The correct answers for all the quiz items were reviewed in the classroom so that all trainees could benefit. The same quiz procedure was followed for the BSF-R motor items. Individuals who continued to have problems were required to attend help labs in the evenings to improve their understanding of the scoring criteria.

After 2 days of direct BSF-R instruction and directed practice role plays, trainees completed a practice exam in preparation for a precertification written exam, which immediately followed the practice exam. The precertification exam included a videotape presentation of a complete BSF-R administration, with core, basal, and ceiling items. Two videotapes showed two testers, each administering the core, basal, and ceiling items of the BSF-R to children. Trainees, as a group, viewed the videotapes item by item and recorded the scores (credit or no credit) they would give for the child's performance on each item. Using a standard BSF-R review form, they also scored the accuracy of the administration of the individual administering the BSF-R on the videotape. The following sample item (exhibit 3-3), which is similar to those on the BSF-R review form, shows that the administration instructions were included so that trainees could compare what the administrator did with the standard instructions. The scoring instructions also were included so that trainees could assign credit or no credit for the child.

Exhibit 3-3. Sample item from the BSF-R review form: 2003–04

3. Uses Means-End Behavior to Retrieve Object

Administration:

During this item, does **administrator**…

1. If using tray table, turn it lengthwise to child?	❑YES	❑NO	❑NA
2. Suspend object and swing 8-10" from child's face at eye level?	❑YES	❑NO	
3. Place object out of reach; string toward child?	❑YES	❑NO	
4. Make any other errors?_____			

Scoring:

During this item, does **child**…

Play with string (doesn't need to grab object)? (Basal Item).....	❑YES	❑NO
Bang object in play? (Basal Item).....	❑YES	❑NO

Score box: Record C/NC in box.	
Uses Means-End Behavior to Grab Object 3.	

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.

Administration of the practice exam gave the trainees time to review their own scoring and receive clarification on any item. The answers for each item were reviewed by the group and questions were discussed. After completing the practice exam, each trainee completed the precertification exam. Although the videotape was paused between items to give trainees time to complete the item, there was no discussion of the answers, and no questions were permitted. Again the trainees scored the child according to the performance in the video (credit/no credit) and critiqued the administration of each item. In order to pass the precertification exam, each trainee had to receive 90 percent or higher on scoring the items and 85 percent or higher on review of the administration. The higher criterion for scoring accuracy was imposed because scoring errors can result in misapplication of the basal or ceiling rule and, therefore, the loss of data. Those trainees who passed the precertification exam were then eligible to attend the live practice certification session.

Any trainee who did not pass either of these precertification criteria was required to attend a mandatory help lab to improve her or his understanding of the administration instructions or the scoring rules before being permitted to advance to the live practice session. In practice, however, almost all trainees voluntarily attended all or some help labs.

The certification process culminated in live practice sessions in which each trainee administered the BSF-R to a child who ranged in age from about 21 to 30 months, the approximate age range trainees would encounter during the ECLS-B data collection. While one trainee administered the BSF-R, the other trainee videotaped the assessment. To ensure that all trainees were obtaining a clearly visible and audible videotape of the BSF-R administration, trainers and technical support staff circulated around the rooms and viewed trainees' camera display screens to make sure that the BSF-R was recorded properly.

During the live practice certification for the BSF-R, lead trainers identified the individuals in their rooms who seemed to be at-risk for not passing. These individuals were live coded by a different lead trainer or field supervisor, with the limitation that a lead trainer or field supervisor could not certify one of his or her "own" trainees or field staff. About 60 trainees were live-coded. The remaining trainees were evaluated from their videotapes. After the live practice sessions, the BSF-R videotapes were reviewed by the ECLS-B training staff.

In order to be certified to administer the BSF-R, each trainee had to earn 90 percent or higher on scoring the child's responses and 85 percent or higher on accuracy of administration. On

average, trainees scored 93 percent for administration accuracy and an average score of 97 percent for scoring accuracy on the BSF-R mental scale, and 96 percent for administration accuracy and 93 percent for scoring accuracy on the BSF-R motor scale. The scores of 20 trainees (out of 135) were considered marginal, and these individuals were targeted for an early quality control visit in the field to monitor their BSF-R administration and scoring. One of this group of trainees resigned. The remaining 19 were reviewed early and scored well on the BSF-R during the quality control home visit.