The ECLS-B Direct Assessment
Choosing the Appropriate Score for Analysis

Cognitive Direct Assessment

9-month & 2-year data waves

The Bayley Short Form – Research Edition (BSF-R). At the base year (henceforth called the 9-month wave, as children were approximately 9 months of age at the time of assessment) and first follow-up (henceforth called the 2-year wave, since children were approximately 2 years of age at the time of assessment), the Bayley Short Form-Research Edition (BSF-R) was used. The BSF-R consists of a subset of items from the Bayley Scales of Infant Development, 2nd Edition (BSID-II). Item Response Theory (IRT) modeling was used to select items from the BSID-II for the BSF-R. The items included on the BSF-R represent all of the constructs covered by the BSID-II, are easy to administer, require few materials, and are straightforward to score. Study children were administered these items (i.e., the BSF-R) in place of the longer BSID-II.

Using IRT modeling, along with information from the 9-month and 2-year BSF-R administration, information from the ECLS-B 18-month field test, and the BSID-II publisher data, the ECLS-B established that performance on the BSF-R could be used to estimate performance on the longer BSID-II. IRT was used to put the BSF-R raw scores on the same metric as the BSID-II and to estimate BSID-II scores. Therefore, all the scale scores on the data file are expressed on the BSID-II metric (i.e., the mental scale ranges from 0 – 178; the motor score ranges from 0 –111).

Developing the BSF-R: The BSF-R was constructed using Item Response Theory (IRT) modeling. IRT uses the difficulty, discriminating ability, and "guess-ability" of each item to place it on a continuous scale. Publisher data were obtained and a two parameter model (difficulty level and discrimination power) was used to evaluate the items; "guess-ability" was not deemed relevant, since infants do not guess. Items at equal intervals of difficulty and with a discrimination power of approximately 1 were chosen and care was taken to be certain that all constructs included in the BSID-II were represented in the BSF-R. Next, the set of possible items was narrowed down by ease of administration and materials needed; items easy to administer and with few materials were chosen over those more difficult to administer and needing more materials. Additionally, scoring needed to be as straight forward as possible, since ECLS-B interviewers varied greatly in their knowledge of child development. Lastly, all things being equal, as many "two-fers" as possible were chosen; these are items that can be scored from one administration (e.g., put 1 block in a cup, put 3 blocks in a cup – the child is given 3 blocks and a cup and both items are scored). The final set of items selected for the BSF-R represented all of the BSID-II constructs. Also, like the BSID-II, the BSF-R has a Mental Scale and a Motor Scale. While the BSID-II has 178 Mental items and 111 Motor items, the BSF-R has 29 Mental items at 9 months and 33 Mental items at 2 years, and 35 Motor items at 9 months and 32 Motor items at 2 years.

Once the final set of BSF-R items was selected, they were organized to approximate the BSID-II age sets. The BSID-II groups items into age sets such that no child is ever subjected to all the items. A child begins with the items in his/her age set (e.g., a 9-month-old would begin with the 9-month age set, which has items appropriate for ages 8-11 months) and, if these are too difficult, the assessor switches to the previous age-set (e.g., the 8-month-old age set, or even the 7-month-old age set). Conversely, if the items are too easy, the child would be administered items from the next age set (e.g., the 10-month-old age set, or even the 11-month-old age set). Knowing when to switch age sets and to which age set to go to requires a great deal of expertise and training. Because interviewers on the ECLS-B were not clinicians, it was necessary to organize the items in such a way as to approximate the age sets, with much clearer rules about when to administer the different sets of items. Consequently, the items chosen for the BSF-R were organized into a Core set (administered to all), a Basal set (administered to those who performed poorly on the Core set), and a Ceiling set (administered to those who performed perfectly or nearly perfectly on the Core set). ECLS-B interviewers were given clear instructions regarding when to administer the Basal and Core sets. For example, in the 9-month wave, if a child got 3 or

fewer items in the Core set correct, they were administered the Basal set. If the child missed 3 or fewer items in the Core set, the Ceiling set was administered. In this way, all children were administered the Core set, and determining when to administer the Basal or Ceiling set was straightforward. Thus, all children were appropriately challenged and assessed.

In addition to the direct assessment component, the BSID-II offers a 30 item Behavior Rating Scale (BRS) to help interpret children's performance on the assessment. For the BSF-R in the ECLS-B, 9 of the 30 BRS items were included for this purpose. These items do not, however, approximate the BRS. They can, however, be used to explain scores that are unusually high or unusually low on the assessment.

Preschool, Kindergarten 2006, and 2007 data waves

Early Reading and Math Assessment batteries. The BSF-R is not developmentally appropriate for preschool aged children. Moreover, by preschool and the advent of formal schooling, the ECLS-B put greater emphasis on understanding academic skills of children before and when they enter formal school settings. Consequently, the BSF-R was replaced with an early reading and math adaptive test battery that was designed specifically for the ECLS-B. In choosing the constructs that were most important for the preschool and kindergarten assessments, aspects of children's development and growth that are key milestones at these ages were considered, as well as the knowledge and skills that are important for school readiness and early school success. Thus, the ECLS-B assessment framework combines a developmental age perspective with a focus on academic curriculum content.

In order to assess these key milestones and early school skills, items for the preschool cognitive assessment battery were drawn from a number of standardized instruments and assessment batteries developed for use in other large-scale studies of preschool-aged children (the Peabody Picture Vocabulary Test (PPVT, various forms), the Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP), the PreLAS® 2000, and the Test of Early Mathematics Ability-3 (TEMA-3). Items used in the Family and Child Experiences Study (FACES), the Head Start Impact Study, and the ECLS-K were also part of the preschool battery. Use of the ECLS-K items in the ECLS-B assessment battery not only took advantage of the instrumentation already developed, but increased the likelihood that the data from the ECLS-B cohort could be statistically compared to ECLS-K scores. For this reason, the majority of the items in the kindergarten 2006/2007 battery come from the ECLS-K, although several preschool items were included in order to link the data waves within the ECLS-B cohort.

The preschool and kindergarten cognitive assessment batteries both begin with two sections from the PreLAS® 2000 (Simon Says, Art Show) and PPVT items. All children are administered these items regardless of home language. Those children who were unable to correctly respond to any of these items were thought to be lacking the necessary English language knowledge to be assessed with an English language assessment. Consequently, they were routed either into a Spanish cognitive battery (if their home language was Spanish) or out of the cognitive battery all together (if they spoke a language other than English or Spanish). All children, however, were administered the psychomotor battery (that replaced the Motor aspect of the BSF-R) regardless of English proficiency. Those children who were able to answer at least one item correctly in the language screener, however, moved directly into the early reading battery. At preschool, this battery was similar to the BSF-R structure in that there was a core set items and second stage sets for those children who either had great difficulty with the core set (basal items administered) or who found the core set below their skill level (ceiling items administered). The kindergarten 2006/2007 early reading battery was structured more like the assessment battery for the ECLS-K in that there was a routing test followed by one of three second stage tests (low, middle, or high). The second stage test given to a child was given depended on routing test performance. After the early reading battery the child was routed into the math battery. For all data waves, the math battery consisted of a routing test followed by one of three second-stage tests (low, middle, and high). The preschool battery also had an assessment of color knowledge, which was administered after the early reading test but before the math battery. No color knowledge assessment was administered in the kindergarten 2006 and 2007 waves.

Cognitive Scores

BSF-R scores. IRT modeling is used to estimate the BSID-II scale score (e.g., true score) from performance on the BSF-R. The BSID-II scale scores are on the file, along with the standard error of these scale scores. Additionally, performance on the BSF-R was used to generate T-scores, a standardized equivalent of the scale score. The ECLS-B also provides proficiency probabilities based on the BSID-II raw scores. Because the BSF-R was equated to the BSID-II, ***no BSF-R scores are on the file.***

Early Reading and Math scores. IRT modeling is used to estimate the early reading and math scale scores (e.g., true scores) from performance on the set of items a child may have been administered (routing and second level tests). These reading (language and literacy) and math scale scores, along with the standard error of these scale scores, are on the data file. Additionally, performance on the reading (language and literacy) and math assessment was used to generate T-scores, a standardized equivalent of the scale score. For the preschool round, IRT modeling was also used to create skill specific cluster scores. These scores estimate the number of items each child would have answered correctly per cluster had all the items been administered.[1] Cluster scores for print familiarity, phonological awareness, and receptive vocabulary from the early reading assessment are on the data file. In addition to cluster scores, there are proficiency probability scores. These provide similar information as cluster scores (information about a specific skill set) but are on a different metric; rather than the number of items correct they denote the probability of being proficient at the skill set and consequently range from 0 -1. In reading, there is a proficiency probability that provides information on letter recognition. In mathematics, there is a proficiency probability that provides information on number and shape recognition. In addition to information on children early reading (language and literacy) and math, there is a color knowledge score. This is non-IRT based score, that reflects children's performance on a brief assessment of their color knowledge during the preschool wave.

Choosing the Appropriate Cognitive Score

Each assessment score provides information on children's cognitive competence in a different way. The choice of the most appropriate score for analysis purposes should be driven by the context in which it is to be used or by the research question.

- **Scale Scores or True Scores** represent the children's performance on the BSID-II or Early Reading (language and literacy) and Math assessment batteries.

    - They are useful in identifying cross-sectional differences among subgroups in overall cognitive development and provide a summary measure of cognitive development useful for correlational analysis with status variables such as demographics, family type, or behavioral measures.

    - The Overall Scale Scores may also be used as longitudinal measures of growth, but it is important to remember that gains made from different points on the score scale have qualitatively different interpretations. For example, at 9-months, Mental scores indicate one's ability to engage in purposeful exploration and babbling, whereas at 2-years, Mental scores indicate one's ability to use expressive vocabulary and count. Additionally, gains can only be evaluated across similar assessments. That is, developmental change between 2-years and preschool cannot be calculated as the difference in 2-year and preschool scale scores because qualitatively different metrics were used in each wave. An appropriate use of the 9-month and 2-year data in understanding the

[1] Due to discontinue rules, some of the children who found the assessment more challenging than others were not administered all of the items in order to prevent unnecessary frustration. However, IRT permits the estimation of these children's performance on these items so that cluster scores can be calculated for all children who received the assessment.

relationships between early development and later achievement would be covariates in analysis models predicting preschool, kindergarten 2006, or 2007 scores.

  - Comparison of gains in scale score points is most meaningful for groups that started with similar initial status. Assuming a Guttman model of skill acquisition, when initial status differs substantially, comparisons of scale score gains may be misleading because the value of the gain score refers to the acquisition of qualitatively different skills.

- **Standardized T-scores** are also overall measures of status at a point in time, but they are norm-referenced. They do not answer the question, "What skills do children have?" but rather "How do they compare with their peers?" The transformation to a familiar metric with a mean of 50 and standard deviation of 10 facilitates comparisons in standard deviation units.

  - T-score means may be used longitudinally to illustrate the increase or decrease in gaps in achievement among subgroups over time, but should not be used to calculate gains (i.e., time 1 should not be subtracted from time 2).

- **Cluster Scores**, although derived form the overall IRT model, provide information about specific skills within a domain.

  - Because each score targets a particular set of skills, they are ideal for studying the details of achievement, rather than the single summary measure provided by the overall scale scores and T-scores.

  - They are useful as longitudinal measures of change because they show not only the extent of gains but also where on the achievement scale the gains are taking place. Thus, they can provide information on differences in skills being learned by different groups, as well as the relationships with processes, both in and out of school, that correlate with learning specific skills.

  - Changes in cluster scores over time may be used to identify the process variables that are effective in promoting achievement gains in specific skills.

- **Proficiency Probability Scores**, similar to cluster scores, provide information about specific skills within a domain. They are presented on a different metric than the cluster scores, but they still reflect children's performance on a group of items, measuring a similar construct, within a similar difficulty point on the assessment. They range from 0-1.

  - Because each proficiency targets a particular narrow set of skills, they are ideal for studying the details of development, rather than the single summary measure provided by the scale scores and T-scores.

  - They are useful as longitudinal measures of change because they show not only the extent of gains, but also where on the development scale the gains are taking place. Thus, they can provide information on differences in skills being acquired by different groups, as well as the relationships with processes that correlate with the acquisition of specific skills.

Socio-emotional Direct Assessment

Videotaping Task

Study children's socio-emotional development was directly assessed during the 9-month, 2-year, and preschool data waves. In order to directly assess the child's socio-emotional development at these time points, it was decided to videotape the child interacting with his/her parent and to code these videotapes/DVDs back in the home office. The parent-child interaction is thought to lay the foundation for all future social interactions, for

it is the first relationship the child experiences. It is this relationship that helps the child to develop social expectations and the ability to regulate emotion during a social exchange. Consequently, assessing this interaction provides a great deal of information about the child's socio-emotional development.

9 months. At 9 months, the Nursing Child Assessment Teaching Scale (NCATS) was used to examine parent-child interactions. The NCATS is part of a larger clinical battery known as the Nursing Child Assessment Satellite Training (NCAST). The parent is asked to choose an activity that the child cannot yet do from a list and then teach the child the activity. This instruction period is videotaped. Whether the child learns the activity is not important.

Trained coders viewed the videotaped interactions and determined whether the parent-child dyad exhibited any of 73 different behaviors (yes or no). The parent's teaching behaviors were coded along with the child's responsiveness. Additionally, the parent's ability to read the baby's cues and respond appropriately was coded. A Total Parent score was computed by adding up the scores on the 50 parent behaviors (yes = 1, no = 0). Higher scores on the Total Parent score indicate greater use of teaching behaviors and greater responsiveness to the baby. A Total Child score was computed by adding up scores on the 23 child behaviors (yes = 1, no = 0). Higher scores on the Total Child score indicate greater clarity of cues and responsiveness to the parent. A Total score is also on the file. The Total score is the sum of all 73 coded behaviors. Higher scores represent dyads that are more responsive to one another and have smoother interactions than dyads with lower scores. While the NCATS does contain subscales, these scales had very low reliabilities (alphas) and, consequently, are not on the ECLS-B file. The item level data for the 73 coded behaviors are on the data file.

2 years and Preschool. For the first follow-up, the decision was made to switch from a teaching task to a play task, since at this age the child is more of an active participant in deciding the nature of the interaction. The child is able to do more than simply respond to the parent's bidding. The Two Bags Task was chosen for use in the ECLS-B. This task was also used in the Early Head Start Impact Study and is a modification of the Three Bags Task used by the NICHD Study of Early Child Care. In order to administer the Two Bags Task, the interviewer gave the parent two bags: one with a book (At 2 years the book was 'Goodnight Gorilla' by Peggy Rathmann and at Preschool the book was 'Corduroy' by Don Freeman) and the other with materials for play (at 2-years this bag contained pots and pans; at preschool it had Play-Doh® and cookie cutters). The parent was instructed to start playing with Bag #1 (the book), but could move to Bag#2 whenever s/he and the child were ready. They had a total of 10 minutes to play. This interaction was videotaped and coded back at the home office.

The Two Bags task yields parent scores and child scores. These scores differ a bit from 2-years to Preschool in order to remain developmentally appropriate. At 2-years, the Two Bags task yields six parent behavior scores (Sensitivity, Positive Regard, Cognitive Stimulation, Negative Regard, Intrusiveness, Detachment) and three child behavior scores (Engagement, Sustained Attention, Negativity). These behaviors are assessed on a 7-point likert scale, with higher scores indicating greater display of the behavior. Additionally, at 2-years one composite score was created: Supportiveness. This composite is the mean of the following scales: Sensitivity, Positive Regard, and Cognitive Stimulation. This composite was created because scores on these scales tend to hang together. Higher scores on Supportiveness indicate more sensitivity, positive regard, and cognitive stimulation (or supportiveness) of the child during the play session.

As mentioned above, at Preschool the Two Bags task yields slightly different scores. Because Sensitivity and Positive Regard were so highly correlated at 2-years, at Preschool, it was replaced with one scale that combines the two constructs: Emotional Supportiveness. Consequently, the 2-year Supportiveness composite is dropped at Preschool. The remained of the parent behavior scores are the same: Cognitive Stimulation, Negative Regard, Intrusiveness, and Detachment. Again there are three child behavior scores, but Sustained Attention is replaced by Quality of Play. Thus, the three child scores are Engagement, Quality of Play and Negativity.

Toddler Attachment Sort – 45 (TAS-45)

In addition to the videotaping task, during the first follow-up or 2-year data collection wave, children's attachment to the parent was assessed using a card sort. While conducting the home visit, field interviewers observed the child's behavior with his/her parent. After leaving the home, the field interviewer used a laptop computer to complete a card sort to indicate what child attachment behaviors were observed. Forty-five cards representing different behaviors known to indicate attachment were first sorted into two categories: "apply" and "not apply," depending on whether or not the child displayed the behavior listed on the card. A second sort then resulted in four piles: "strongly apply," "mildly apply," "mildly not apply," and "strongly not apply." These four piles form a 4-point likert scale. Examples of behaviors found on the cards are "Relaxes when in contact with mother," and "Is happy to be alone without mother." The results of these card sorts were then analyzed using multi-dimensional scoring.

In multi-dimensional scoring, the data are examined in the context of a staple "map." In the case of attachment, nine fixed points or hot-spots, each corresponding to a different construct (e.g. Warm and cuddly) on a three-dimensional graph are used to evaluate the data. Each of the 45 items in the card sort has a place on the map. Those items that help to define a construct or hot spot are located closest to the construct's fixed point (i.e., hot spot). So for example, the item "child often hugs or cuddles against parent without being asked to do so" is located close to the hot spot "warm and cuddly" on the three dimensional map. Each of the items that contribute to a construct or hot-spot is given a weight indicating how much that item contributes to that construct. The sum of these weights is the value of that fixed point, or hot spot. Thus, each child's score on the item (the 1-4 Likert scale mentioned above) is multiplied by the corresponding weight and then summed up to get a value on the hot spot. The hot spots have a potential range from -2 to 2. Negative values mean that the child did not really evidence that behavior or characteristic during the field interviewer's visit. Positive values suggest that the interviewer did observe evidence of that characteristic while in the home. For example, a negative value on Warm and Cuddly means that the field interviewer did not see the child acting in warm and cuddly ways toward the parent. Conversely, a positive value suggests the opposite; the child did show warm cuddly behavior toward the parent while the interviewer was present.

Next, the "map" and the Security and Dependency scores are used to determine the child's attachment classification: Insecure Avoidant (A), Secure (B), Insecure Ambivalent (C), and Disorganized (D). The hot-spots are plotted on a horizontal line. Each child's scores are plotted (e.g., child#1 is high on Warm and Cuddly (+ value), low on Avoids Others (negative value) and so on. These plots are then connected to form a best fit line. All the children are plotted on the same graph. The securely attached children, those with a B classification stand out from all the others; their best fit lines cluster together in a distinct pattern. Next, children's data are plotted on a two dimensional graph with one axis representing Security and the other representing Dependency. Where the insecurely attached children fall on this graph is used to distinguish the insecure attachment classifications (i.e., A, C, & D) from one another. Thus, all children of a particular attachment classification share a similar profile, though some may be more positive or more negative on certain characteristics or hot spots. It should be noted though, that the conditions of the home visit are different from a classic Strange Situation. The child is in the comfort of his/her own home and may not be stressed to the same degree as during a Strange Situation.

Socio-emotional Scores

There are two types of Socio-emotional scores on the ECLS-B data file: videotaping scores and attachment scores.

- **Videotape task scores:**
  - Total Parent Score (NCATS)
  - Total Child Score (NCATS)
  - Total Score (NCATS)
  - Parental Sensitivity (Two Bags, 2-years)

- Parental Positive Regard (Two Bags, 2-years)
- Parental Emotional Support (Two Bags, Preschool)
- Parental Cognitive Stimulation (Two Bags)
- Parental Negative Regard (Two Bags)
- Parental Intrusiveness (Two Bags)
- Parental Detachment (Two Bags)
- Parental Supportiveness composite (Two Bags, 2-years)
- Child Engagement (Two Bags)
- Child Sustained Attention (Two Bags, 2-years)
- Child Quality of Play (Two Bags, Preschool)
- Child Negativity (Two Bags)

- **Attachment scores:**
  - Attachment Classification
  - AQS Security & Dependency scores
  - Hotspot Scores

Choosing the Appropriate Socio-emotional Score

- **Videotape task scores** provide information on the child's early interaction experiences. Analysts may be interested in determining whether different groups of parents interact with their children in different ways and the effects these differences may have on the child's development. Thus, scores yielded from the videotaping task may be used either as predictor or outcome variables.

- **Attachment classification** permits analysts to examine children by their particular attachment style. This is the most well known and widely used attachment measure and is sufficient to examine the association of various predictors with children's status of attachment as an outcome variable.

- **AQS Security & Dependency scores** permit those analysts familiar with the AQS security and dependency scores to use these scores as they would those from the AQS, rather than the attachment classification. The security score indicates the child's ability to use the adult as a secure base. The dependency score is an indication of clinginess to the parent. Both of these scores range from – 1 to 1. Scores closer to 1 on security indicate a high ability to use the parent as a secure base, whereas scores closer to -1 indicate a low ability to use the parent as a secure base. Likewise, scores closer to 1 on dependency indicate high clinginess, whereas scores closer to -1 indicate low clinginess. These scores can each be used on their own to examine concurrent associations, for example, between security and exploratory competence, or between dependency and fearfulness. Longitudinal associations can also be examined, for example, between security and school achievement or between dependency and adjustment to school. Analysts interested in using these scores should note that –1 is a valid score for both these variables so that the conventional use of reserve codes in the data file does not apply to these variables.

- **Hotspot Scores** represent different characteristics of the child's attachment behavior: Warm and Cuddly, Cooperative, Enjoys Company, Independent, Attention Seeker, Upset by Separation, Avoids Others, Demanding/Angry, and Moody/Unusual. The child's score on any one hotspot is derived by first multiplying the child's score on an item by its corresponding weight and then summing up. For example, if a child got a 4 on the item "Cooperates with mother and gives her things if asked" then that score of 4 would be multiplied by the weight given to that item with respect to the Cooperative hot spot. All the items for the Cooperative hot spot would be treated in this manner and then summed up for a Cooperative hot spot score. All hot spot scores have a possible range from -2 to +2. Positive scores indicate more cooperativeness whereas negative scores represent less cooperativeness. It is recommended, however, that analysts *use the attachment classifications* when examining issues related to attachment. The hotspot values on the data file are ordinal. Strictly speaking, they are proportionate in nature and not of the

"classic" Likert-type. Researchers who are not interested in investigating attachment, per se, would be able to use children's scores on a hotspot to examine associations between the various hotspot domains and children's outcome measures. For example, a researcher interested in exploring the development of children's social competence could examine associations between the hotspots or characteristics and measures of social functioning in subsequent data collections.[2]

Physical Direct Assessment

The ECLS-B, at each data wave, assesses the child's height (length at 9 months), weight, Middle Upper Arm Circumference (MUAC), and for those children born of very low birth weight, head circumference. Additionally, at 2-years, Preschool, and Kindergarten 2006 child's Body Mass Index (BMI) was calculated. Measurement protocols were consistent with those on the NHANES.

Physical Measure Scores

There are several physical measure scores on the data file:

- **Height/Length (cm)**
- **Weight (kg)**
- **MUAC (cm)**
- **Head Circumference (cm)**
- **Body Mass Index (BMI)**

As is standard practice in all major health studies, each of these physical measurements was obtained at least twice. Thus, there are two measurements for each of the above in the data file. An average of these two measurements was then taken to get the child's measurement at that point in time. The longitudinal 9-month–Preschool data file includes the composite X3CHBMI, which indicates children's body mass index (BMI) and is calculated using the formula: (weight in kilograms/(height in centimeters)2) * 10,000. Similar scores will be created for the kindergarten 2006 and kindergarten 2007 datafiles.

[2] For more information on the development of Attachment hotspot scores, please see Kirkland, J., Bimler, D., Drawneek, A., McKim, M, and Schölmerich, A. (2004). An alternative approach for the analyses and interpretation of attachment sort items. *Early Child Development and Care, 174(7-8),* 701-719.