# Early Childhood Longitudinal Study, Birth Cohort (ECLS-B)

## Psychometric Report for the 2-Year Data Collection

**U.S. Department of Education**
Institute of Education Sciences
NCES 2007–084

## Methodology Report

**August 2007**

Carol Andreassen
Philip Fletcher
**Westat**

Jennifer Park
*Project Officer*
**National Center for Education Statistics**

**U.S. Department of Education**
Margaret Spellings
*Secretary*

**Institute of Education Sciences**
Grover J. Whitehurst
*Director*

**National Center for Education Statistics**
Mark Schneider
*Commissioner*

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high-priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high-quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public. Unless specifically noted, all information contained herein is in the public domain.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

August 2007

The NCES World Wide Web Home Page address is http://nces.ed.gov.
The NCES World Wide Web Electronic Catalog is http://nces.ed.gov/pubsearch.

This publication is only available online. To download, view, and print the report as a PDF file, go to the NCES World Wide Web Electronic Catalog address shown above.

**Suggested Citation:**
Andreassen, C., and Fletcher, P. (2007). *Early Childhood Longitudinal Study, Birth Cohort (ECLS–B) Psychometric Report for the 2-Year Data Collection* (NCES 2007–084)*.* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

**Content Contact:**
Gail Mulligan
(202) 502-7491
gail.mulligan@ed.gov

## ACKNOWLEDGMENTS

Over the past 7 years, many individuals and organizations have contributed to the design and conduct of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B). While it is not possible to name all the individuals who have made significant contributions to this study, we would like to recognize some of those who played a critical role during the development and implementation phases of the ECLS-B.

First, we would like to thank all the children and their parents who participated the study thus far. The parents of these children invited us into their homes and allowed us to work with their children. We would also like to thank the more than 200 field staff and 16 field managers and supervisors who conducted the home visits during both national data collections and the several earlier field tests of the study design and instrumentation. Also, we would like to express our appreciation to the State Vital Registration and Statistics Executives who provided the sample of 2001 birth certificates on which the study is based.

We gratefully acknowledge Marian MacDorman of the National Center for Health Statistics (NCHS) for all her efforts to make the ECLS-B a success. Her tireless work on the project's behalf to help gain the cooperation of the states and her review of the many sets of review documents submitted to the NCHS and state institutional review boards (IRBs) were invaluable.

Westat, under the direction of the National Center for Education Statistics (NCES), designed and conducted the 2-year wave of the study. NCES received day-to-day support from the Education Statistics Services Institute (ESSI), and American Institutes for Research (AIR). NCES would also like to thank Rick Dulaney and Ray Saunders of Westat for the leadership they provided to the programming, systems development, and data processing tasks associated with the study; Richard Hilpert of Westat for overseeing the field operations of the study; and Laura Branden of Westat for her tireless work with the NCHS and state IRBs.

The design of the child assessment battery also benefited from the participation of expert work groups. The design of the Bayley Short Form–Research Edition was guided by experts in child assessment and the Bayley Scales of Infant Development, Second Edition (BSID-II) in particular, and in Item Response Theory. These individuals included Don Rock and Judy Pollack of Educational Testing

Service who provided Item Response Theory analysis expertise; Barbara Wasik of the Johns Hopkins University who advised on child assessment with low-socioeconomic status and language-minority children; and Kathleen Matula who had been involved in the restandardization of the BSID-II for The Psychological Corporation and who advised on item administration and scoring and training of field staff. John Kirkland of Massey University, Palmerston North, New Zealand, and his colleagues, David Bimler and Andrew Drawneek, were vital to the creation of the TAS-45 and the development of the laptop application. Rachel Cohen (Administration on Children, Youth and Families, U.S. Department of Health and Human Services), Jeanne Brooks Gunn and Christy Brady-Smith of Teachers College were consulted about the administration and coding of the Two Bags Task and generously provided a special coding training for Westat's trainers. In addition, Howard Hoffman (National Institute on Deafness and Other Communication Disorders, National Institutes of Health) worked with Philip Dale (University of Missouri) to provide the wordlists for the English and Spanish vocabulary assessments.

We also would like to acknowledge Jerry West, formerly of NCES and currently of Mathematica Policy Research, Inc., whose dedication and extraordinary efforts brought the ECLS-B into being, as well as Jennifer Park (NCES), Kristin Denton Flanagan (AIR), Jodi Jacobson Chernoff (AIR), and Sandy Eyster (AIR), whose efforts on the ECLS-B are reflected in the study instrumentation, data processing activities, and data products. In addition, the authors would like to thank Zeyu Xu, Steve Mistler, and Rob Stillwell of AIR, as well as Chris Chapman, Larry Ogle, Emmanuel Sikali, Ralph Lee, Val Plisko, and Marilyn Seastrom of NCES for their thoughtful reviews of this document.

## Sponsoring Agencies

## Early Childhood Longitudinal Study, Birth Cohort (ECLS-B)

- National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education (ED)
- National Center for Health Statistics, U.S. Department of Health and Human Services (HHS)
- National Institutes of Health (NIH), U.S. Department of Health and Human Services

  National Institute of Child Health and Human Development
  National Institute on Aging
  National Institute of Mental Health
  National Institute of Nursing Research
  National Institute on Deafness and Other Communication Disorders
  Office of Behavioral and Social Sciences Research
  National Center on Minority Health and Health Disparities

- Economic Research Service, U.S. Department of Agriculture (USDA)
- Administration on Children, Youth and Families, HHS
- Maternal and Child Health Bureau, Health Resources and Services Administration, HHS
- Office of Special Education Programs, ED
- Office of the Assistant Secretary for Planning and Evaluation, HHS
- Office of Indian Education, ED
- Centers for Disease Control and Prevention, HHS
- Office of Minority Health, HHS

*This page is intentionally left blank.*

## CONTENTS

## CONTENTS (continued)

## CONTENTS (continued)

Chapter Page

## CONTENTS (continued)

## LIST OF TABLES

## LIST OF TABLES (continued)

## LIST OF TABLES (continued)

## LIST OF TABLES (continued)

## LIST OF TABLES (continued)

## LIST OF FIGURES

## LIST OF FIGURES (continued)

## LIST OF FIGURES (continued)

## LIST OF FIGURES (continued)

## LIST OF FIGURES (continued)

# LIST OF FIGURES (continued)

## LIST OF EXHIBITS

# 1. INTRODUCTION

The Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) is a multisource, multimethod study that focuses on the early home and educational experiences of young children. The study is following a nationally representative cohort of children born in 2001 from birth until they enter kindergarten. The first round of the study was conducted when the children were approximately 9 months old; the present phase of the study, round 2, consists of data collected when the children were approximately 2 years of age. Three additional rounds of data collection are planned: one at preschool (about 4 years of age) and two at kindergarten entry.[1]

The ECLS-B includes direct and indirect assessments of children's competencies and skills to provide data on their developmental status at a given point in time, as well as growth over time. This report documents the design, construction, implementation, quality control, and psychometric characteristics of the direct and indirect child assessments in the second round of data collection. Direct assessments, as the name suggests, are obtained by directly administering assessments to the children and scoring the results. The indirect assessments are parent respondent or caregiver respondent[2] ratings of children's characteristics, abilities, or behaviors. This report parallels the *ECLS-B Methodology Report for the Nine-Month Data Collection, Volume 1: Psychometric Characteristics* (NCES 2005–100) (Andreassen and Fletcher 2005). After presenting a brief overview of the ECLS-B and the data collection instruments used in round 2, this chapter discusses the following topics:

- Literature reviews conducted to inform the study design;
- The role of the Technical Review Panel (TRP) in identifying appropriate developmental domains and associated instruments for research and in ensuring the collection of high quality data from the field;
- The child assessment working group;
- The original ECLS-B design and the results of field testing, with subsequent redesign; and
- The organization of the remainder of the report.

[1] There will be two kindergarten rounds of data collection – one in fall 2006 and one in fall 2007. In the fall of 2006, data will be collected from *all* participating sample children, about 75 percent of whom are expected to be enrolled in kindergarten. In the fall of 2007, data will be collected only from the children who are entering kindergarten for the first time, which is expected to be about 25 percent of the sample children.

[2] The person who provided this information was usually, but not always, the child's parent.

## 1.1 Overview of the ECLS-B

The ECLS-B is part of the Early Childhood Longitudinal Study (ECLS) program, a longitudinal studies program comprising two cohorts—a birth cohort of children born in 2001 (ECLS-B), and a kindergarten cohort of children who were kindergartners in the 1998-99 school year (ECLS-K). The goal of the ECLS program is to provide high quality data on children's development and growth in the early childhood years that are useful for researchers, policymakers, practitioners, and parents. Together, the ECLS-B and the ECLS-K provide the breadth and depth of data required to more fully describe children's health, early learning, development, and education experiences. See http://nces.ed.gov/ecls for information about the ECLS program.

The central goal of the ECLS–B is to provide a comprehensive and reliable set of data that may be used to describe and to better understand children's early development; their health care, nutrition, and physical well-being; their preparation for school; key transitions during the early childhood years; their experiences in early care and education programs and kindergarten; and how their early experiences relate to their later development, learning, and experiences in school. The design of the ECLS-B was guided by three principles. The first 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.

The first round of data for the ECLS-B was collected when the children born in 2001 were approximately 9 months old. In that round, about 10,700 children and their parents participated; about 9,850 of these children and their parents participated in the second round of data collection when the children were approximately 2 years old. Child assessments were conducted on a total of about 10,200 of these children at 9 months and on about 9,200 of these children at 2 years. Of these 9,200 children, about 8,950 have assessment data for both rounds of data collection. Details on the sample design, sample selection, and data collection can be found in the *ECLS-B Methodology Report for the Nine-Month Data Collection, Volume 2: Sampling* (NCES 2005–147) (Bethel et al. 2005) and in the *User's Manual for the ECLS-B Longitudinal 9-Month–2-Year Data File and Electronic Codebook* (NCES 2007-046) (Nord et al. 2006). Users who want more detailed information about the 9-month data collection should consult the *ECLS-B User's Manual for the Nine-Month Restricted-Use Data File and Electronic Codebook* (NCES 2004-092) (Nord et al. 2004).

The ECLS-B is sponsored by the U.S. Department of Education, National Center for Education Statistics (NCES) in the Institute of Education Sciences, in collaboration with several health, education, and human services agencies, including the National Center for Health Statistics (NCHS), the National Institutes of Health (NIH), the Administration on Children, Youth and Families (ACYF), and the U.S. Department of Agriculture (USDA). Westat, a social science research organization, conducted the first two rounds of the study for NCES.

## 1.2 Data Collection Instruments for the 2-Year Data Collection

The ECLS-B 2-year data collection took place between January 2003 and April 2004. Data were collected by computer-assisted personal interviews (CAPI) with parent respondents,[3] self-administered questionnaires given to parent respondents and resident and nonresident fathers (if appropriate),[4] direct child assessments during an in-person home visit, and from field staff observation of the children's behavior and home setting during the home visit. For children with regular child care arrangements, data were also obtained by computer-assisted telephone interview (CATI) with the child care provider; for a randomly selected sample of children in child care, a direct observation of the child care setting was conducted and, if the setting was center-based, a self-administered questionnaire was given to the director of the center. Exhibit 1-1 lists the sources of data in the 2-year data collection.

[3] The respondent to the parent CAPI was usually, but not always, the child's parent. For 1.1 percent of the 9,850 round 2 cases, the respondent was a non-parent relative, such as a grandparent, or a nonrelative.

[4] If the resident father was not present during the home visit, the father questionnaire was left with the respondent to give to the father to complete. If the father was present, the interviewer gave the questionnaire directly to him. Most resident fathers were the child's biological, adoptive, step-, or foster father. However, for fewer than 100 cases, the Resident Father Questionnaire was completed by someone other than the child's father (e.g., by a grandfather). Interviewers sought the mothers' permission to contact the nonresident father for an interview. If the nonresident father was present, permission still had to be obtained from the mothers before giving the Nonresident Father Questionnaire to him. Only biological fathers were contacted to complete the nonresident father questionnaire.

Exhibit 1-1. Sources of data and instruments in the ECLS-B 2-year data collection: 2003–04

| Instruments |
|---|
| Parent computer-assisted personal interviewing (CAPI) Instrument |
| Direct child assessments (using CAPI, paper and pencil, and videotapes) |
| Parent Self-Administered Questionnaire |
| Resident Father Self-Administered Questionnaire |
| Nonresident Father Self-Administered Questionnaire |
| Child Observations and Interviewer Remarks Questionnaire |
| Child Care Provider telephone interview |
| Child Care Observation for a subset of the children interviewed;<br>Center Director Self-Administered Questionnaire for center directors |

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.

The ECLS-B 2-year direct child assessments consisted of four components: the Bayley Short Form–Research Edition (BSF-R), the Two Bags Task, the Toddler Attachment Sort (TAS-45), and physical measurements. Exhibit 1-2 displays the major domains measured during the direct child assessments by each component. Interviewers administered the components using a hard-copy booklet called the Child Activity Booklet, which was available in both English and Spanish. BSF-R item scores and physical measurements were recorded in the Child Activity Booklet. The instructions for the Two Bags Task were also included in the Child Activity Booklet.

Exhibit 1-2. Components and substantive domains covered in the ECLS-B 2-year direct child assessments: 2003–04

| Child assessment component | Domain coverage |
|---|---|
| Bayley Short Form–Research Edition (BSF-R) | Cognitive (mental), physical (motor) |
| Two Bags Task | Socioemotional functioning, cognitive stimulation fostered by parent, and child's engagement with parent and willingness to learn |
| Toddler Attachment Sort | Security of attachment |
| Physical measurements (height, weight, middle upper arm circumference, head circumference[1]) | Physical growth and development |

[1] Head circumference was measured only for ECLS-B sampled children who were very low birth weight.
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.

The 2-year data collection includes the same three direct assessments of children's developmental status, socioemotional functioning, and physical growth and development that were included in the 9-month data collection. It is important to recognize that the ECLS-B assessments do not provide information on children's global mental ability or intelligence quotient (IQ). Instead, they provide descriptive information on children's early cognitive, motor, and social competencies, which are skills that are important for school readiness and early school success. Assessment in the ECLS-B serves three purposes: (1) to describe children's developmental status at particular time points, (2) to examine growth in children's development over time, and (3) to explore the relationship of early experiences to children's development (where assessment data are the outcome).

## 1.3 Literature Reviews

To plan the design of the ECLS-B, three literature reviews were prepared for NCES and are available as working papers on the NCES website at http://nces.ed.gov/ecls/. These working papers included: (1) *Formulating a Design for the ECLS: Review of Longitudinal Studies* (NCES Working Paper Series, Working Paper No. 97–24) (Green et al. 1997); (2) *A Birth Cohort Study: Conceptual and Design Considerations and Rationale* (NCES Working Paper Series, Working Paper No 1999–01) (Moore et al. 1999); and (3) *Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children* (NCES Working Paper Series, Working Paper No. 96–18) (Meisels, Atkins-Burnett, and

Nicholson 1996). Please refer to chapter 2 of the *User's Manual for the ECLS-B Nine-Month Restricted-Use Data File and Electronic Code Book* (NCES 2004–092) (Nord et al. 2004) for further details about these literature reviews. In addition, Child Trends (a nonprofit research firm located in Washington, DC) reviewed available measures in eight domains considered important for the ECLS-B, including the child assessments (Moore et al. 1999). This review of available measures was particularly important for the design of the indirect and direct child assessments because it summarized specific assessments of children's cognitive and psychomotor development, socioemotional functioning, and physical growth and development. This review was the starting point for the design of the child assessment portion of the ECLS-B. With the scope of the child assessment outlined by these literature reviews, Westat then gathered information to guide the selection of the specific measures to be used.

In addition to the above literature reviews, Westat staff reviewed the questionnaires and child assessments that have been used in comparable large-scale, nationally representative studies to determine their operational feasibility for inclusion in the ECLS-B. At the same time, Westat staff reviewed published direct assessments of children's developmental status (e.g., the Bayley Neurodevelopmental Screener, the Bayley Scales of Infant Development, Second Edition (BSID-II), the Mullen Scales of Early Learning, and the Denver Developmental Screening Test) to evaluate whether any of these were operationally feasible for administration by interviewers in the ECLS-B field setting. The main emphasis was to identify the most likely candidates for a standardized measure of children's developmental status in the ECLS-B and evaluate their respective psychometric and administrative strengths and weaknesses.

In recent decades, the literature on parent-child interaction has accumulated reliable findings of positive associations between maternal sensitivity and responsivity and children's developmental outcomes. Methodologies described in the literature, as well as in the above-cited literature reviews, were examined to identify feasible methods for obtaining accurate, reliable, and valid information about aspects of parenting and parent-child interaction known to be predictive of children's later adjustment to, and achievement in, formal schooling. This review of available measures of parenting and parent-child interaction also was used to identify emerging areas of interest in developmental and educational psychology to ensure that the ECLS-B included assessments that were relevant to future research needs. For example, in developmental psychology, self-regulation is increasingly regarded as an important aspect of temperament and, therefore, a set of questions was included in the parent interview to obtain information about infant's self-regulatory behaviors (Aksan and Kochanska 2004; Kochanska, Coy, and Murray 2001; Raver 2004; Bornstein and Suess 2000).

## 1.4 Technical Review Panel

The design of the content of the child assessment component of the ECLS-B has been guided by a Technical Review Panel (TRP). The TRP is a panel of advisors from the research, policymaking, and educational communities who contribute to the ECLS-B by ensuring that it meets the diverse needs of the represented groups. As expert advisors and reviewers, the TRP members help to ensure the success of the ECLS-B in a number of ways, including commenting on overall research priorities, and reviewing and commenting on technical issues. These issues include designing and implementing the ECLS-B; providing information about emerging policy and research topics appropriate for the ECLS-B to address; reviewing questionnaires and assessment instrument content; reviewing draft reports; and reviewing operational practices. An important responsibility of the TRP is to ensure that the plans for conducting the ECLS-B are well thought out and complete, and this responsibility requires a broad range of expertise.

TRP members reviewed the quality of both the design plans and the data collection procedures for the child assessments, and discussed these plans and procedures, as well as alternatives, at the TRP meetings. The members performed the following tasks:

- Verified that the chosen assessments addressed aspects of child development that were determined to be integral to the ECLS-B purpose;
- Assessed whether the chosen instruments were reliable and valid measures of the constructs they were intended to measure;
- Introduced emerging policy issues and research topics to ensure that the information needed to address them was being collected; and
- Reviewed the plans for collecting the child assessments to make sure that their implementation would ensure the collection of high quality data and valid results.

During the early design phases of the 9-month data collection, the TRP meetings took place twice a year in Washington, DC. The TRP members met in a plenary session on the first day, along with representatives from NCES, interagency partners, and Westat staff. After the plenary group discussed general issues, the TRP members then divided into four smaller work groups that covered four content areas: (1) maternal and child health; (2) cognitive and language development and home environment; (3) socioemotional development; and (4) the family’s community, father involvement, and child care. The TRP work groups then reported back to the plenary group, and their comments, suggestions, and recommendations were discussed and taken into consideration by NCES, the interagency partners, and Westat.

For the 2-year collection, the TRP, NCES, and interagency partners felt that direct assessment was an important and unique feature of the study and should be continued. With respect to the domains that should be assessed, all were in agreement that children’s language use was an important milestone to capture. During children’s second year, their language use is burgeoning. Language acquisition is a critical developmental milestone because it is the foundation for mental representation and continued cognitive growth. Language acquisition is also sensitive to environmental and experiential influences, such as mother-child interaction and family literacy. Therefore, it was determined that it was important to include strong measures of toddlers’ developmental status, language abilities, socioemotional functioning, physical growth, and well-being, which are most accurately measured through direct assessment of the child.

To measure developmental status, the consensus was to continue to use both the mental and motor scales of the shortened version of the BSID-II, called the Bayley Short Form–Research Edition (BSF-R) that was used in the 9-month collection. The advisors also recommended that children’s socioemotional functioning continue to be assessed with an observational measure of parent-child interaction. Noting that the construct of security of attachment is a key developmental milestone for children at this particular age, the advisors recommended that a direct assessment of children’s attachment status be included. Given the design of the 2-year ECLS-B home visit, the prime candidate for assessing attachment status was the Attachment Q-Sort (Waters and Deane 1985). However, this tool is too complex for administration in the field by interviewers untrained in attachment theory. (Please see chapter 8 for a discussion of attachment theory.) Therefore, at the suggestion of the TRP members, a simplified and shortened version of the Attachment Q-Sort was developed, the Toddler Attachment Sort-45 item (TAS-45), as described in chapter 8.

A recommendation was also made for the direct assessment of children’s physical growth and well-being using standard measurements of physical growth commonly used in health studies, such as the National Health and Nutrition Examination Survey (NHANES; further information is available online at http://www.cdc.gov/nchs/nhanes.htm), as well as middle upper arm circumference (MUAC), which is used in health studies conducted by the World Health Organization (WHO). These standard measurements include children’s height, weight, MUAC and, for very low birthweight babies (i.e., 1,500 grams or less), head circumference. In addition, since 2-year-old children are able to stand on a scale independently, it was, therefore, possible to assess body mass index (BMI), a commonly used measure of body fat based on height and weight that applies to individuals 2 years of age and older. BMI for children, also referred to as BMI-for-age, is sex and age specific. BMI-for-age is plotted on sex-specific growth charts. These charts are used for children and teens 2 to 20 years of age. For the 2000 Centers for Disease Control and Prevention (CDC) Growth Charts, please refer to the CDC's NCHS website at http://www.cdc.gov/growthcharts/.

The recommended indirect child assessments in the ECLS-B obtained two types of information. The first type was a continuation of the parent-reported assessments of children's developmental functioning used at 9 months, such as the subset of items from the Infant/Toddler Symptom Checklist (DeGangi et al. 1995) and questions about developmental milestones, making age-appropriate modifications. A checklist of words spoken by the child and level of grammar use was also added, to be completed as part of the parent computer-assisted personal interview (CAPI) instrument. This checklist obtained information about the child's language use and complexity of syntax. The second type of indirect measurement obtained information about the child's home experiences and home environment. This information was obtained through interviewer observations and questions asked of the parent respondent.

## 1.5 Specialized Child Assessment Work Group and Expert Consultants

As design work progressed, a specialized assessment work group was formed to consult on the development of the 2-year BSF-R. In addition, appropriate experts were consulted to guide decisionmaking and the design and implementation of other measures including the Two Bags Task and the TAS-45. The guidance of the assessment work group is discussed in chapter 2. Expert guidance for the development of the coding workshop for the coding of the Two Bags Task is discussed in chapter 6. The development of the TAS-45 is discussed in chapter 8.

## 1.6 Design Change: 18-Month and 30-Month Data Collections and Transition to the 2-Year Data Collection

Originally there were to be two data collections during the toddler-to-preschool period: one at 18 months and one at 30 months. Planning for the direct assessments, therefore, focused on the development of an 18-month BSF-R and on a 30-month BSF-R. The plan for the direct assessments at 18 months was to develop an age-appropriate version of the BSF-R (analogous to the 9-month BSF-R) to obtain measures of children's cognitive and psychomotor development. The Nursing Child Assessment

Teaching Scale (NCATS), a semi-structured parent-child interaction activity that had been used at 9 months, would again be used as an observational measure of characteristics of the mother-child relationship. The physical measurements would be obtained again using the same procedures as at 9 months.

At the 30-month data collection, an age-appropriate version of the BSF-R would be developed to obtain measures of children's cognitive and psychomotor development. The observational measurement of mother-child interaction would be changed to the Three Bags Task in order to build a foundation for continuity of measurement with the preschool observational measure. Because continuity of measurement was an important guiding principle for the design of the ECLS-B, the plan was to use the NCATS at 9 and 18 months and the Three Bags Task at 30 months and preschool so that the same measure would be used at two data collection points. By 30 months, BMI could also be obtained.

While the 18-month field test was in progress, design work on the 30-month BSF-R and Three Bags Task was conducted. Near the end of the 18-month field test, NCES and the interagency partners decided to combine the 18- and 30-month data collections into a single data collection at 2 years. At this point, the 18-month field test, which included the BSF-R, had been completed. Also, the 30-month BSF-R had been designed and the items pilot tested by child development staff at Westat. In addition, testing the format of a 30-month BSF-R administration booklet and the Three Bags Task had begun. The pilot test of the 30-month BSF-R items gave Westat child development staff direct experience with the items. They could then eliminate items that were not feasible for administration in the field.

When the decision was made to combine the data collections into a single 2-year collection, Westat was able to take the results of the 18-month field test and the 30-month pilot work to produce an age-appropriate 2-year BSF-R and to select age-appropriate mother-child activities for the Three Bags Task at 2 years. The NCATS was not appropriate for 2-year olds so it was not carried into the 2-year design process. The Three Bags Task was selected as an age-appropriate 2-year old mother-child activity. The pilot test of the Three Bags Task for the 30-month data collection indicated that it would take too much time to administer, so a recommendation was made to reduce it to two activities; the Three Bags Task became the Two Bags Task. In addition, the transition to a 2-year data collection also meant that children's BMI could be obtained as discussed earlier. These design changes are discussed in greater detail in chapter 2 for the BSF-R, chapter 6 for the Two Bags Task, and chapter 7 for the physical measurements.

## 1.7 Organization of This Report

Subsequent chapters describe the specific assessment instruments and sets of questions included in the 2-year ECLS-B data collection and summarize how each performed in the field. Chapter 2 presents a discussion of the decision to include a direct assessment of children's developmental status, the BSID-II, and the adaptation of the 18-month and 30-month shortened versions of that assessment for use at 2 years. Chapters 3 and 4 describe the work that was done to develop the BSF-R and the Item Response Theory (IRT) analyses conducted with the longitudinal dataset (i.e., combined 9-month and 2-year BSF-R scores). Chapter 5 describes the BSF-R scores included on the longitudinal 9-month–2-year data file. The Two Bags Task—the observational measure of the videotaped parent-child interaction—is summarized in chapter 6, and children's physical measurements are summarized in chapter 7. Chapter 8 discusses the work that was done to develop the TAS-45—a shortened version of the Attachment Q-Sort—and summarizes scores obtained, and the procedures for training interviewers and maintaining reliability during the year of data collection. Chapter 9 summarizes the remaining observation items that were completed by the interviewer in CAPI after the home visit was completed, including the interviewer observations of child behavior during the BSF-R, and the child's home environment. Chapter 10 summarizes the indirect assessments of the child in the parent CAPI instrument, including the toddler word list, developmental milestones, and children's self-regulatory skills. Finally, a table of the direct child assessment intercorrelations is presented in appendix A and the Toddler Attachment Sort items are presented in appendix B. A list of references is provided at the end of this document. Throughout this report, a brief review of key features of the 9-month assessments relevant to the 2-year measures is included when warranted.

Please note that two longitudinal weights that can be found on the longitudinal 9-month–2-year data file were used to obtain the estimates reported in this document. Please see section 4.5.1.2 of the user's manual for more information on these weights. These two weights are W2R0 and W2C0. Weight W2R0 represents cases with completed parent interviews at both rounds 1 and 2. It was used to estimate child-level characteristics associated with data collected through the parent interview or birth certificate, or both. Examples relevant to child assessment include sets of questions in the parent interview addressing children's ages when developmental milestones were reached, children's self-regulation behaviors, and children's home environments. Weight W2C0 represents cases with completed parent interviews and completed child assessments at both rounds 1 and 2. It was used to estimate child-level characteristics associated with data collected through the child assessments—either alone or in combination with data collected through the parent interview or birth certificate, or both. Examples include children's scores on the direct assessment of cognitive functioning and psychomotor functioning, children's and primary caregivers' scores on the observational measurement of socioemotional functioning, and children's physical measurements.

# 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.

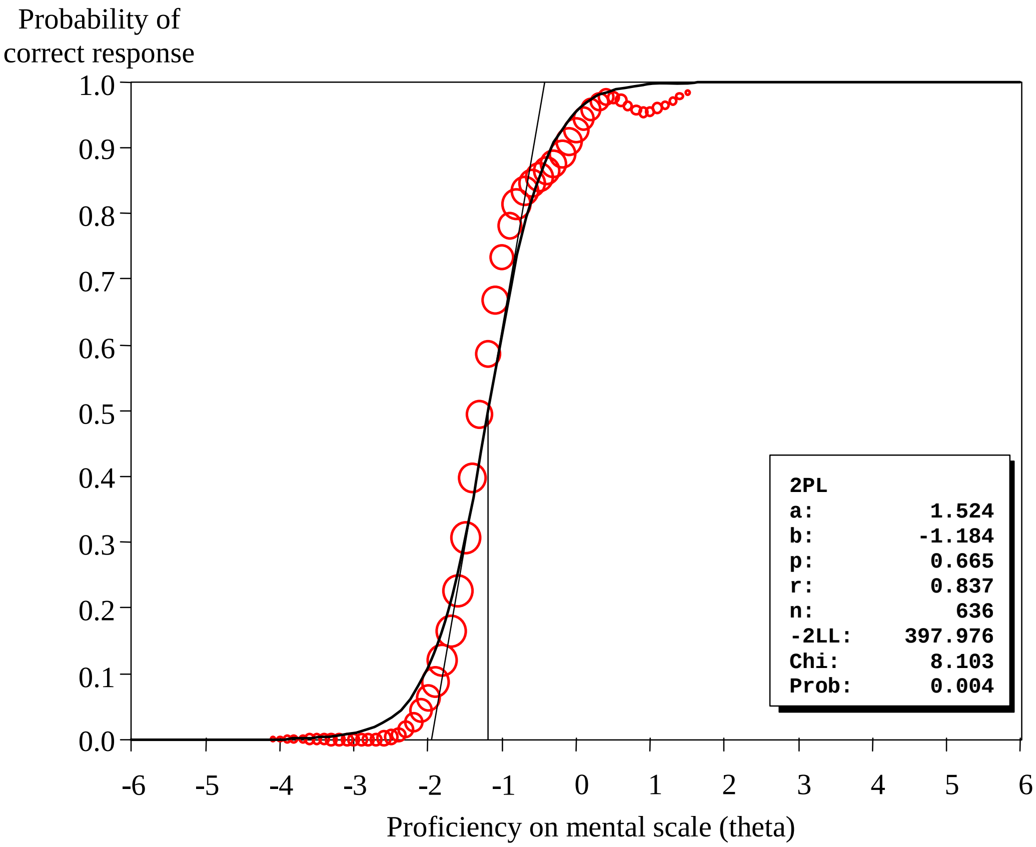
### 2.1.3 Overview of the 2-PL Response Model

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 $\theta$ 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 $[-\infty < \theta < \infty]$.

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.

The formula for the 2-PL response function is:

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

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 $\theta$; 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 $\theta$ and item difficulty $b_i$. The greater the value of item difficulty $b_i$, the greater the ability $\theta$ 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 population[2] 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 discriminating[4] 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 $\theta$ 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

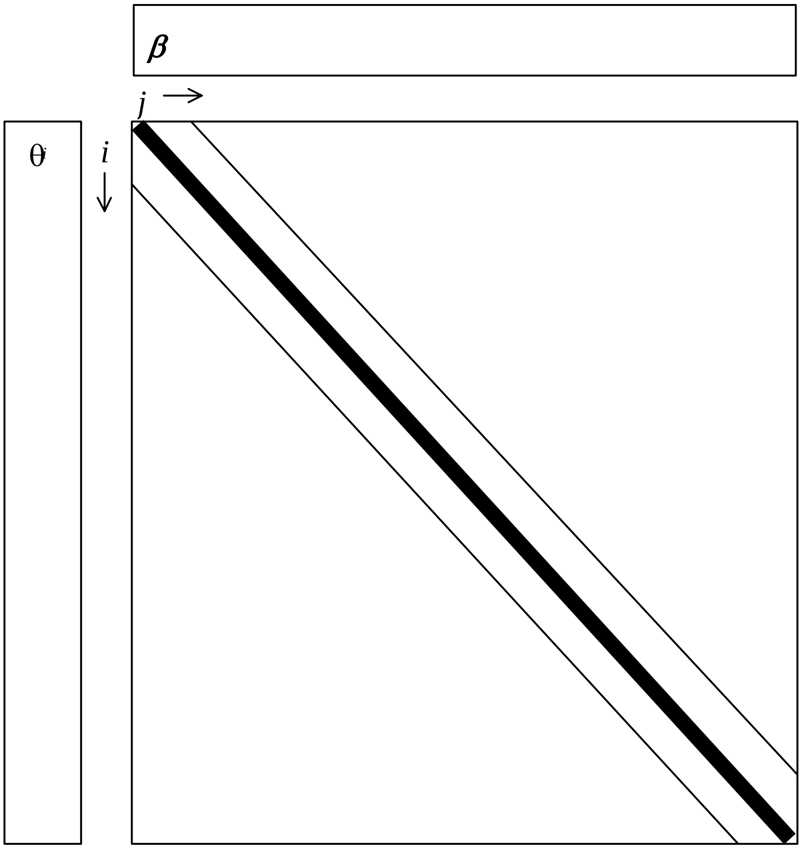
| | | Mental scale | | | | Motor scale | | | |
|---|---|---|---|---|---|---|---|---|---|
| Months | Sample | Raw score | | Index score | | Raw score | | Index score | |
| of age | *N* | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 1 | 100 | 15.3 | 9.4 | 101.8 | 18.2 | 11.7 | 3.9 | 101.5 | 13.5 |
| 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 $\theta_i$, and item parameters $\beta_j$: 1993



NOTE: $i$ = rows of individuals sorted by individual ability ($\theta_i$) and $j$ = columns of items sorted by item difficulty ($\beta_j$).
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.

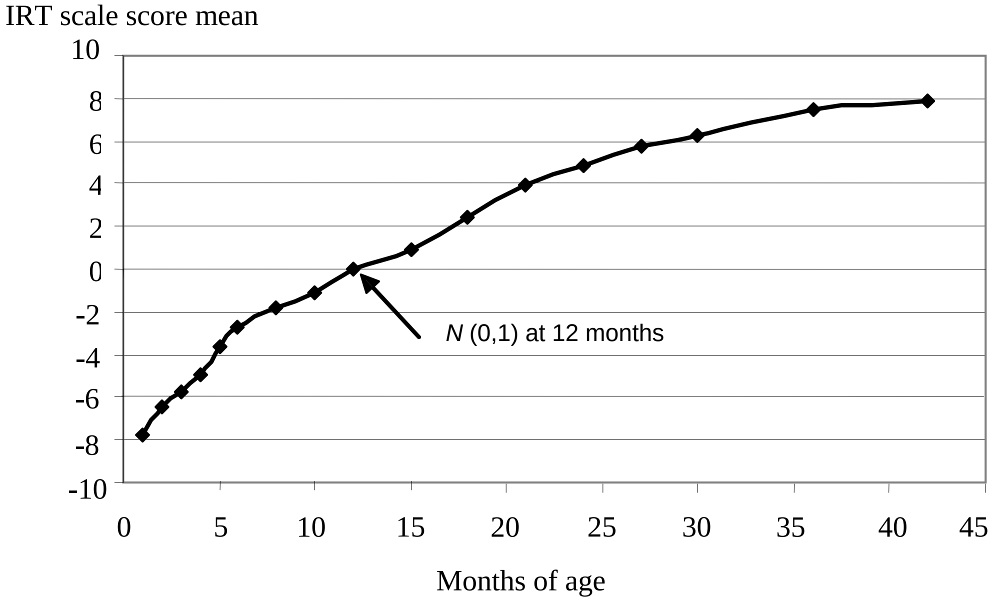
Common item linkages are used to calibrate the full set of BSID-II items on the mental and motor scales spanning development between 1 and 42 months of age. Item calibrations require that a latent population distribution be chosen to establish an IRT metric for ability and difficulty parameters. The 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.

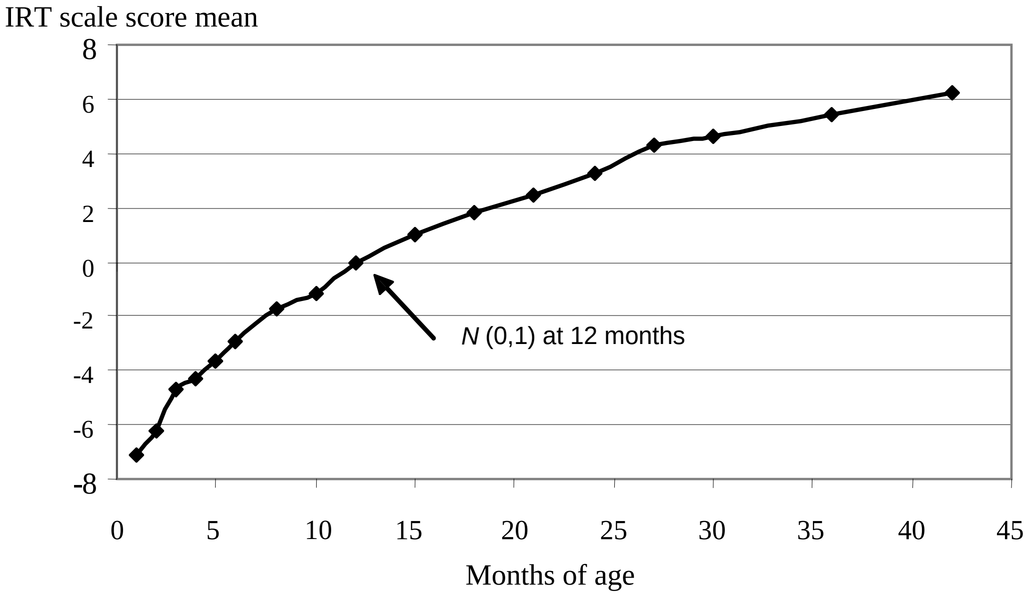
Figure 2-2. BSID-II mental scale score means by age: Item Response Theory (IRT) 2-parameter logistic item calibrations using publisher data: 1993



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.

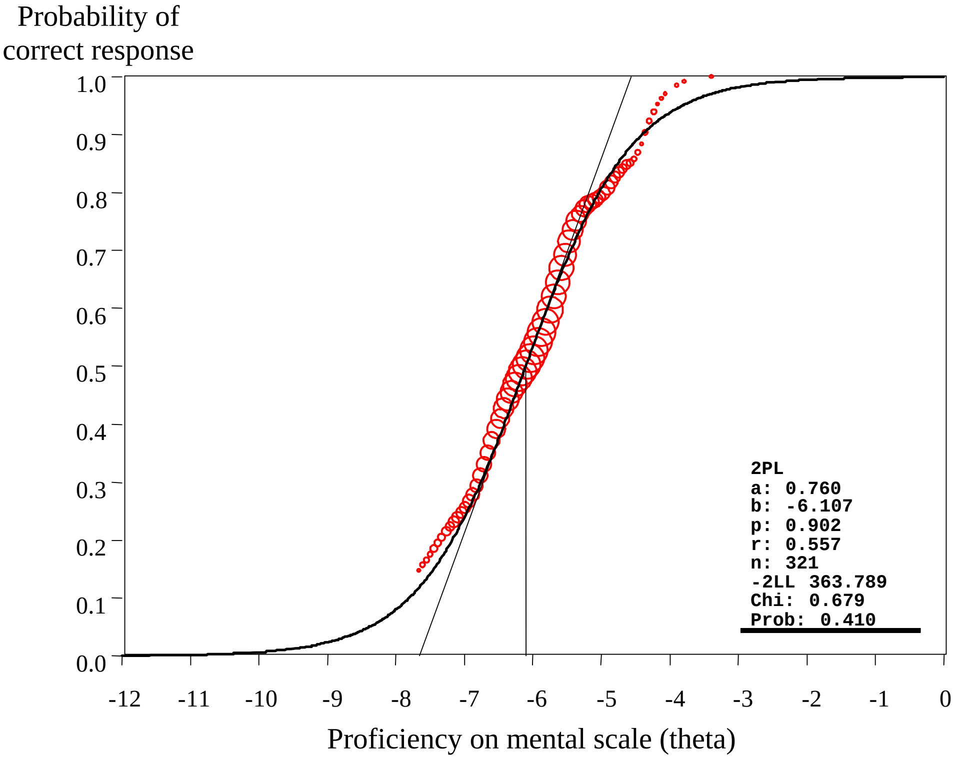
Figure 2-3. BSID-II motor scale score means by age: Item Response Theory (IRT) 2-parameter logistic item calibrations using publisher data: 1993



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 $\chi^2$ 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 $\theta$. 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.

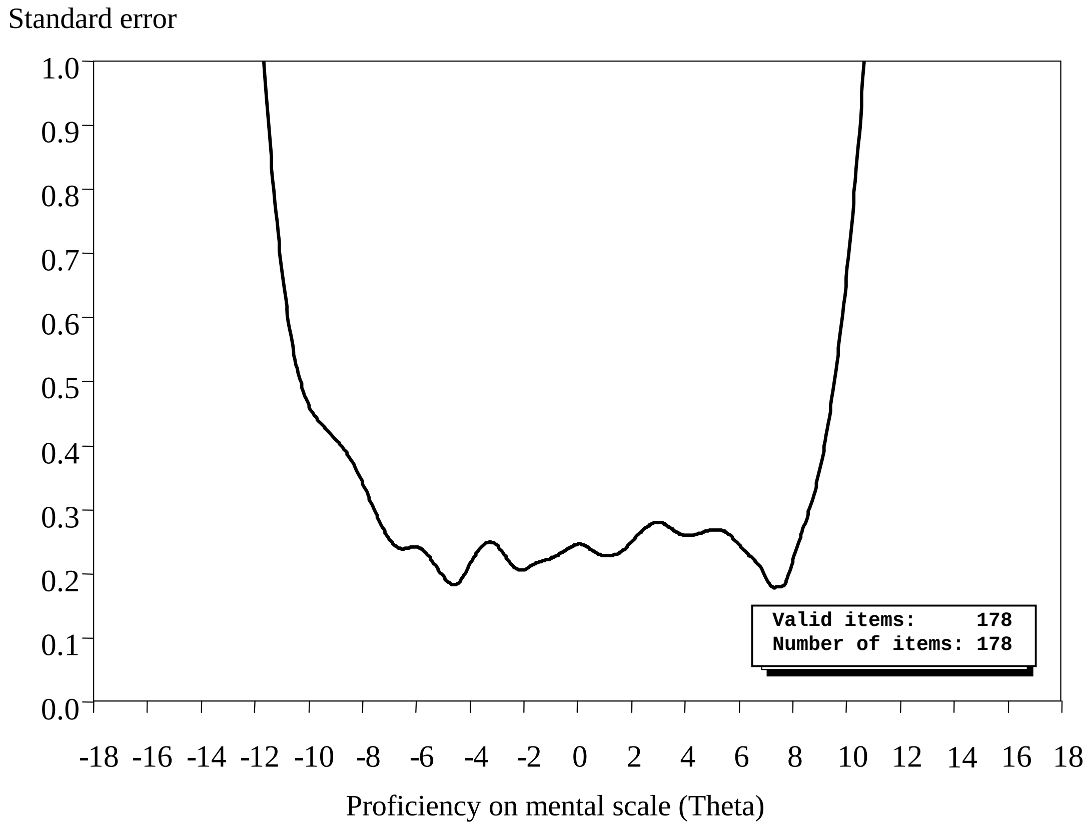
Collectively, the 178 mental and 111 motor items convey an extraordinary amount of information about children. The items are numerous, discriminate well, and are age appropriate in relation to the target population. These conditions produce tests that are both reliable and informative. High levels of 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 1937[5]) 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.

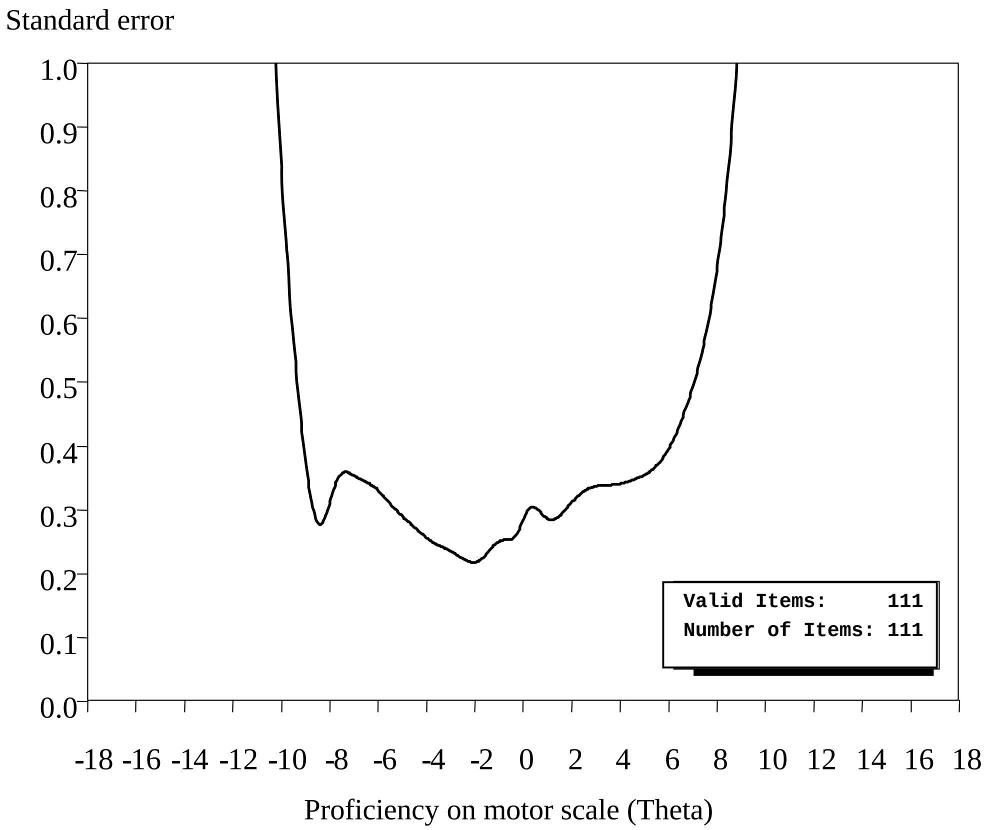
[6] These coefficients are IRT equivalents of KR-20 coefficients. Although similar to coefficient alpha, the more general symbol for reliability, $r_{xx}$, is used.

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

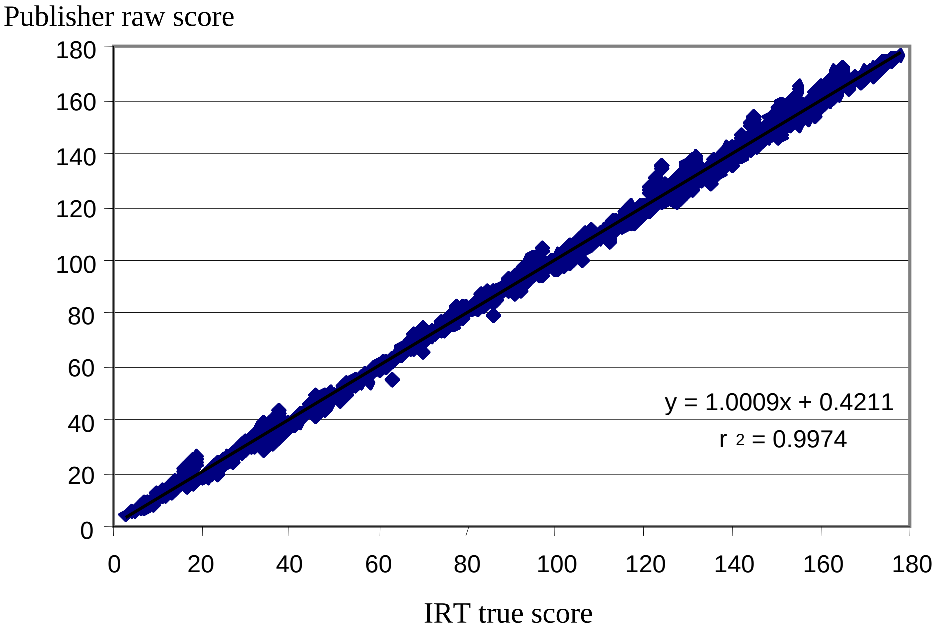


SOURCE: Standardization dataset for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

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 $\theta$ 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 $\theta$: $\xi = \sum_{j=1}^{n} P_j(\theta)$. 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.

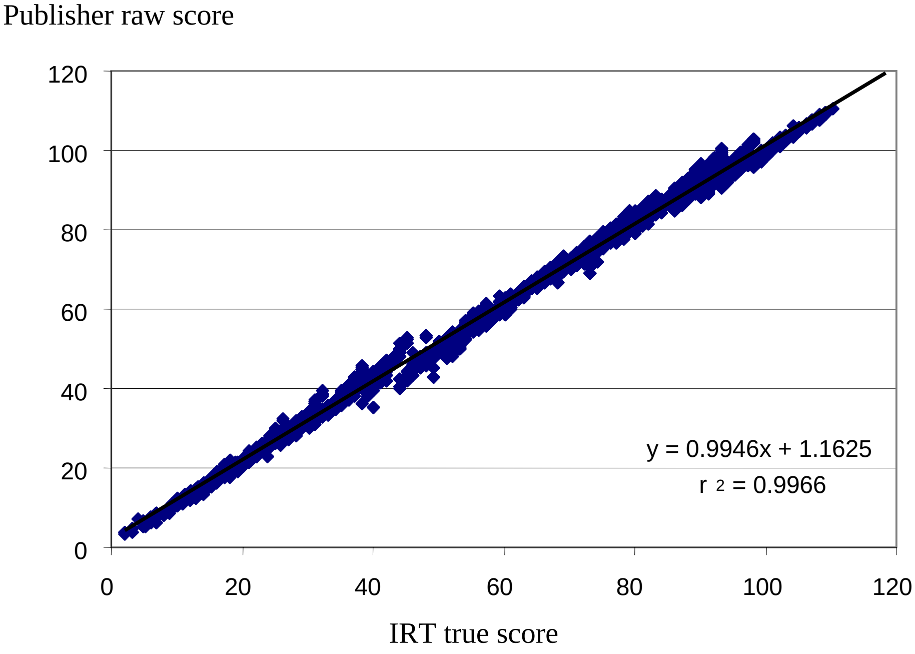
Figure 2-7. Relationship between Item Response Theory (IRT) true score and publisher raw score for the mental scale: IRT 2-parameter logistic item calibrations using publisher data: 1993



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.

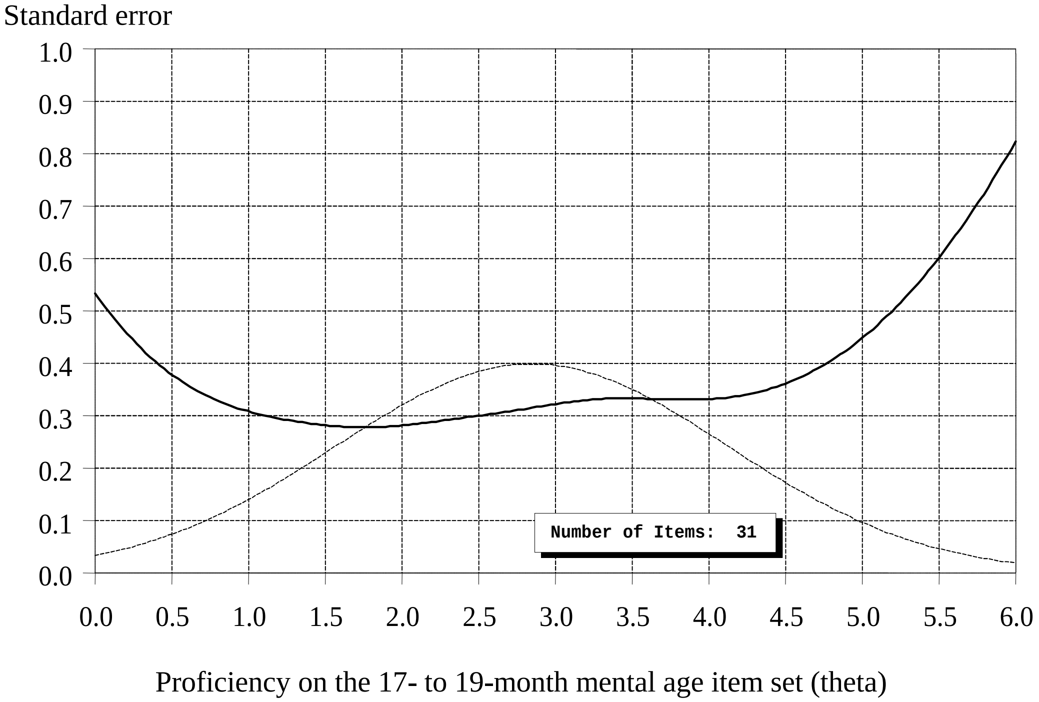
In order to select reduced item sets for the 18-month BSF-R, the technical properties (i.e., difficulty and discrimination parameters) of items in the 18-month[8] age item set recommended by the publisher were examined. There are 100 observations for this age group in the publisher standardization sample.

### 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.

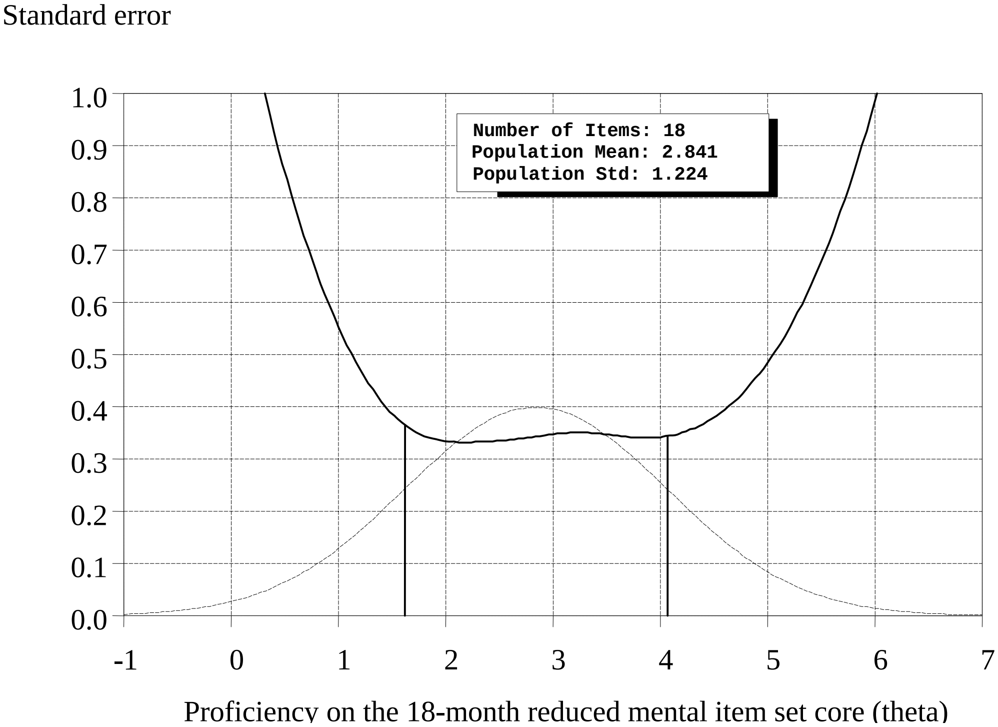
The standard error depicted in the figure shows that the 31 items in the age item set recommended by the publisher afford considerable measurement precision for 18-month-old children within the limits of $\mu \pm \sigma$. Moving outward from the mean, growth in the standard error of measurement accelerates, 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



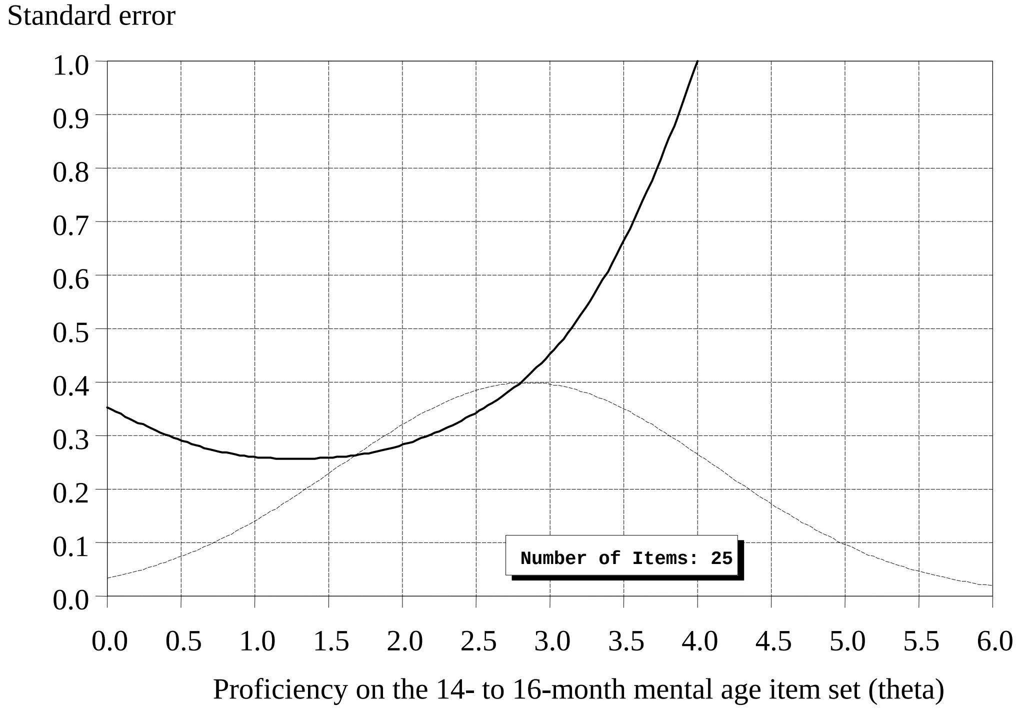
NOTE: The solid lines descending from the curve to the x-axis indicate 1 standard deviation above and 1 standard deviation below the mean. Std = standard deviation. The normal curve (dashed line) represents the projected latent distribution of the 18-month population and is included for illustration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

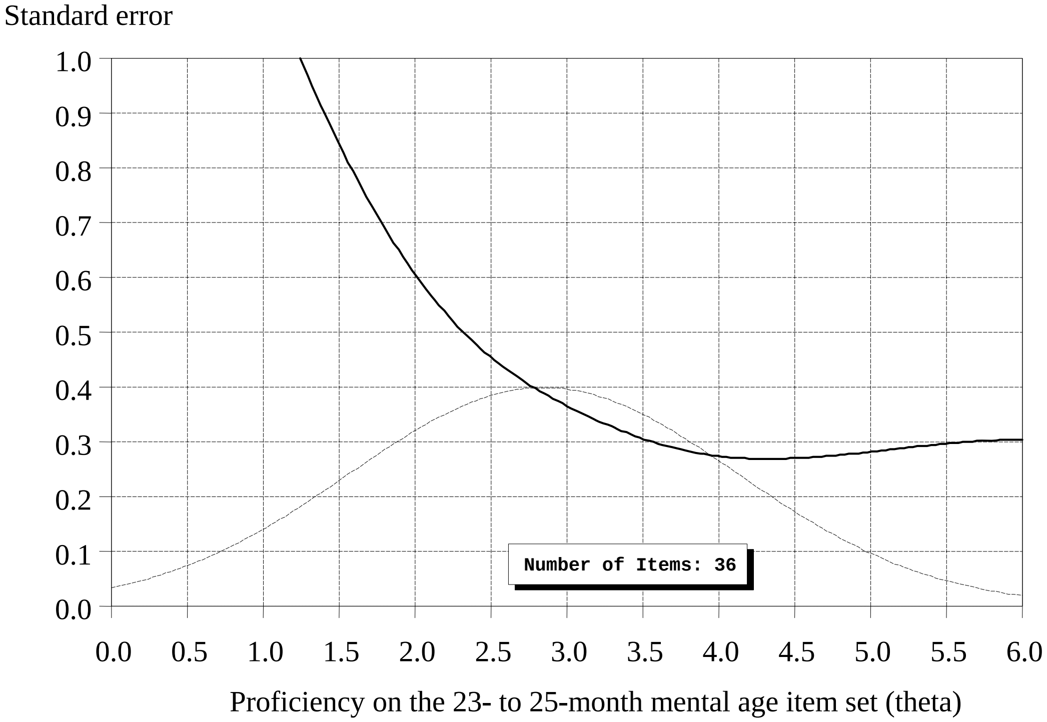
Figure 2-11. Standard error of measurement by proficiency level for the BSID-II mental scale 14- to 16-month age item set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



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 μ + $\sigma$. 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

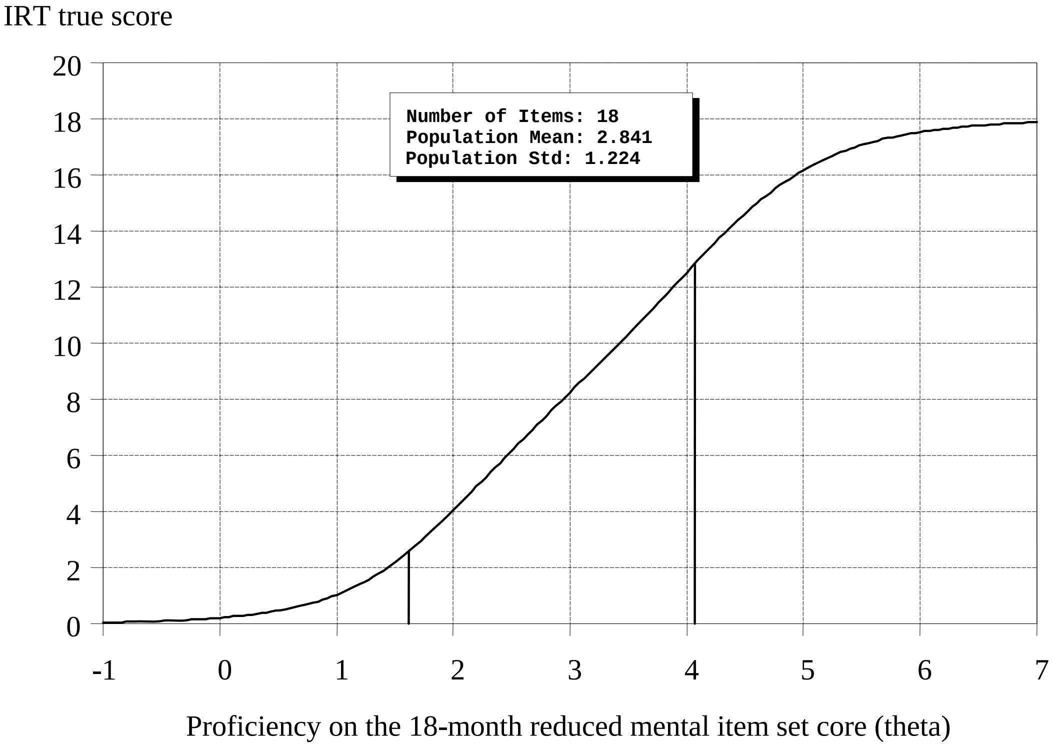


NOTE: The publisher does not provide data for all months of age in the calibration dataset but does provide recommended items for different months of age. The normal curve (dashed line) represents the projected latent distribution of the 23- to 25-month population and is included for illustration purposes.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

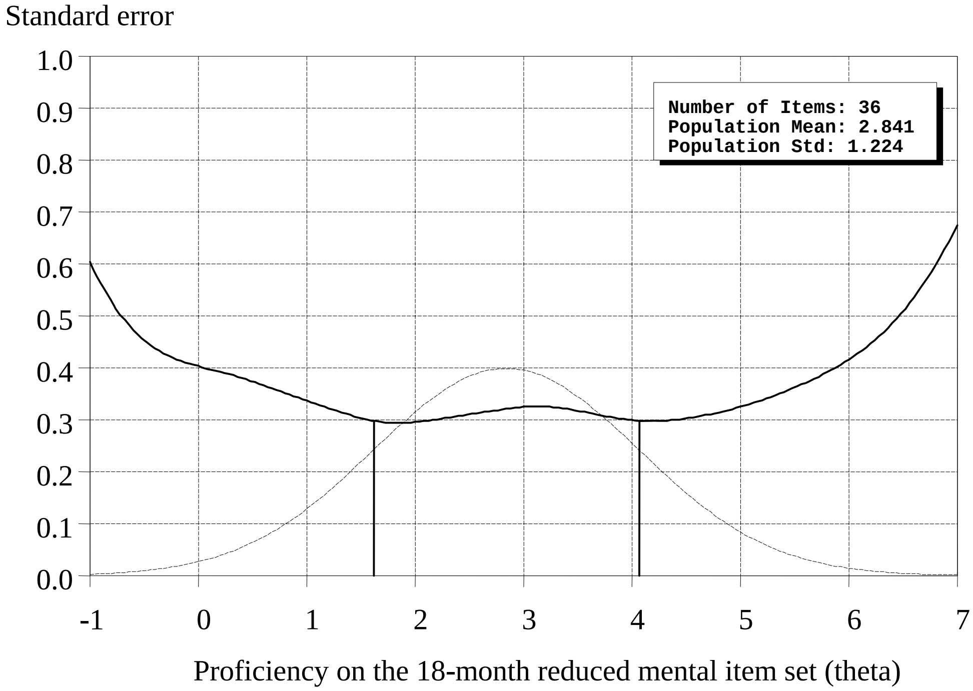
Reading the true score value opposite each of the vertical lines at the point where they join the curve provides an estimate of the number-right score at each of these limits. The values are approximately 3 at the lower end and 13 at the high end of this range. The conservative decision rule that was 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 percent x 18) + (16 percent x 26) + (16 percent x 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 |

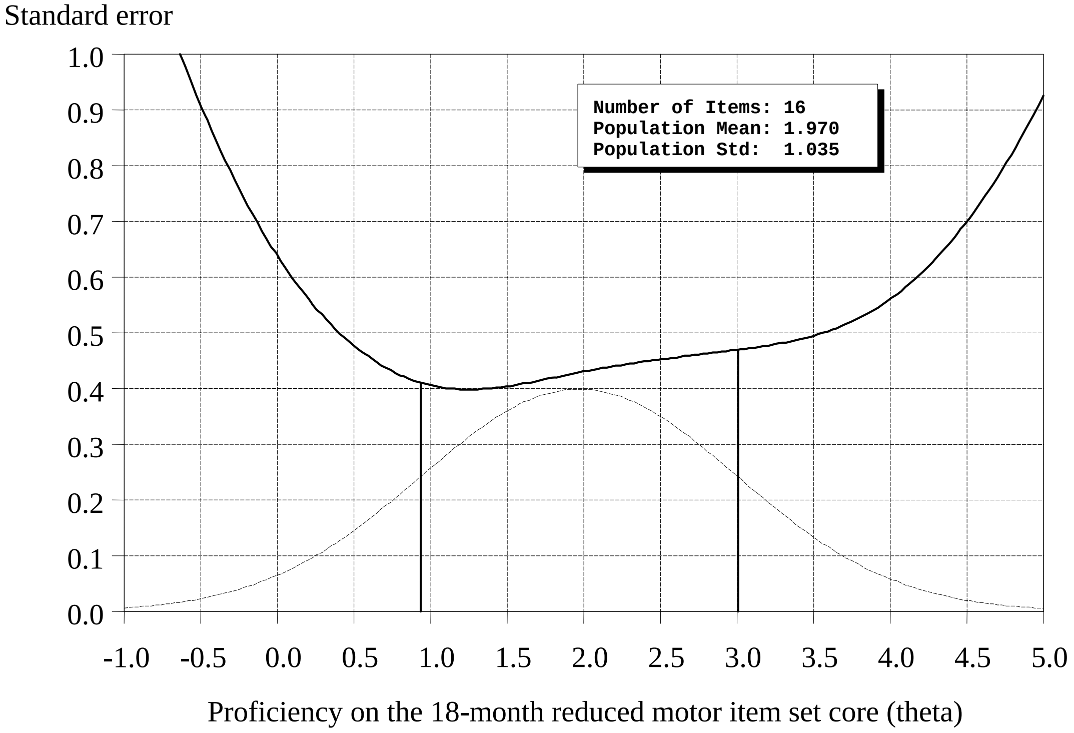
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 2-year data collection, 2003–04.

### 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

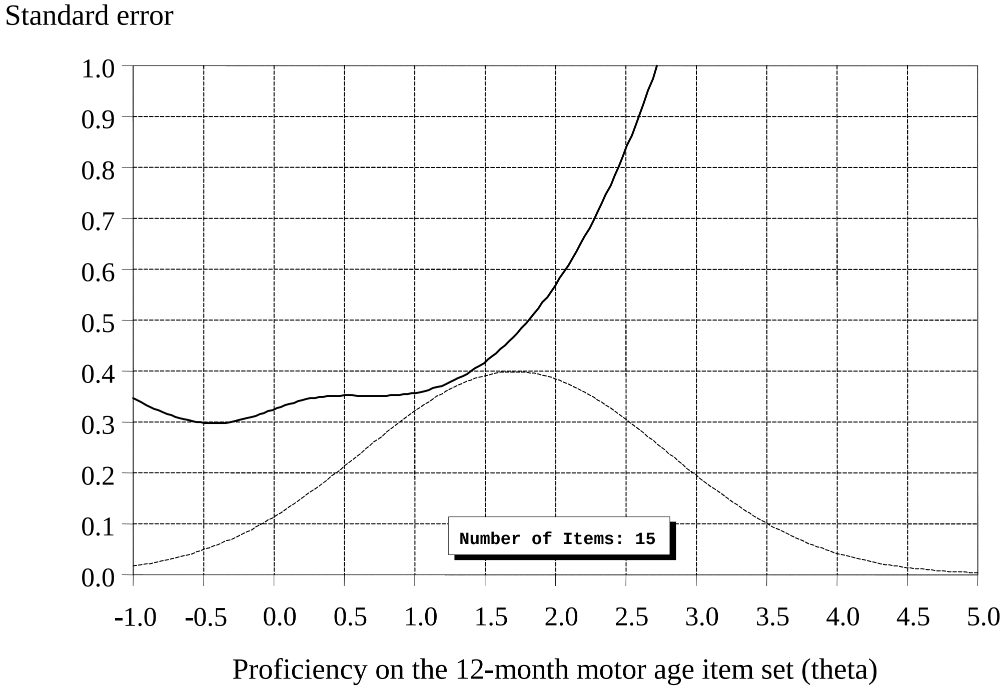


NOTE: The solid lines descending from the curve to the x-axis indicate 1 standard deviation above and 1 standard deviation below the mean. Std = standard deviation. The normal curve (dashed line) represents the projected latent distribution of the 18-month population and is included for illustration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

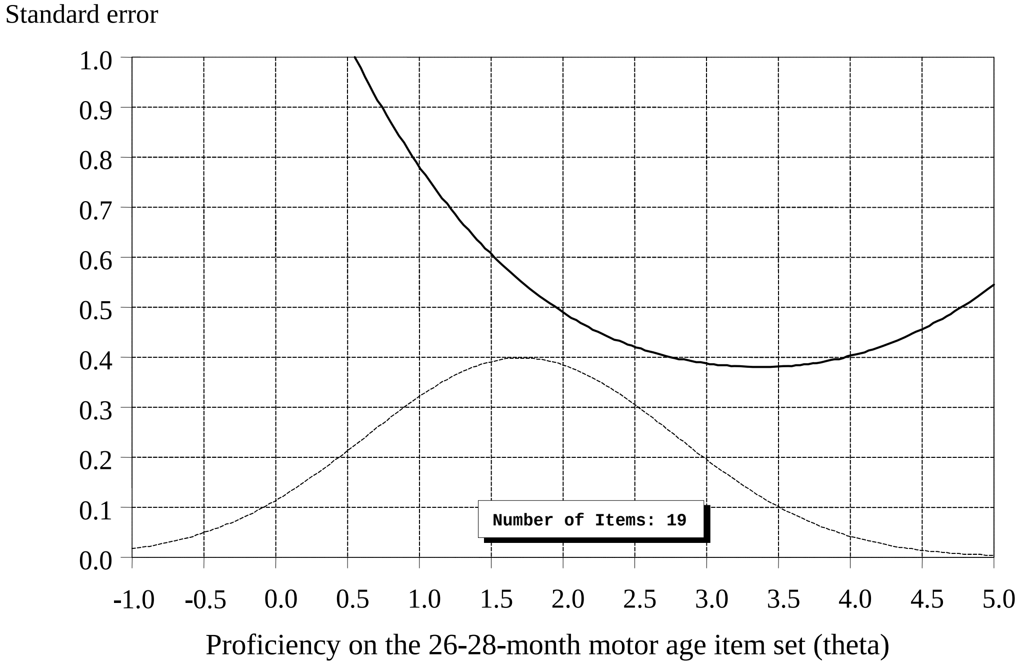
Figure 2-16. Standard error of measurement by proficiency level for the BSID-II 12-month motor age item set: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



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



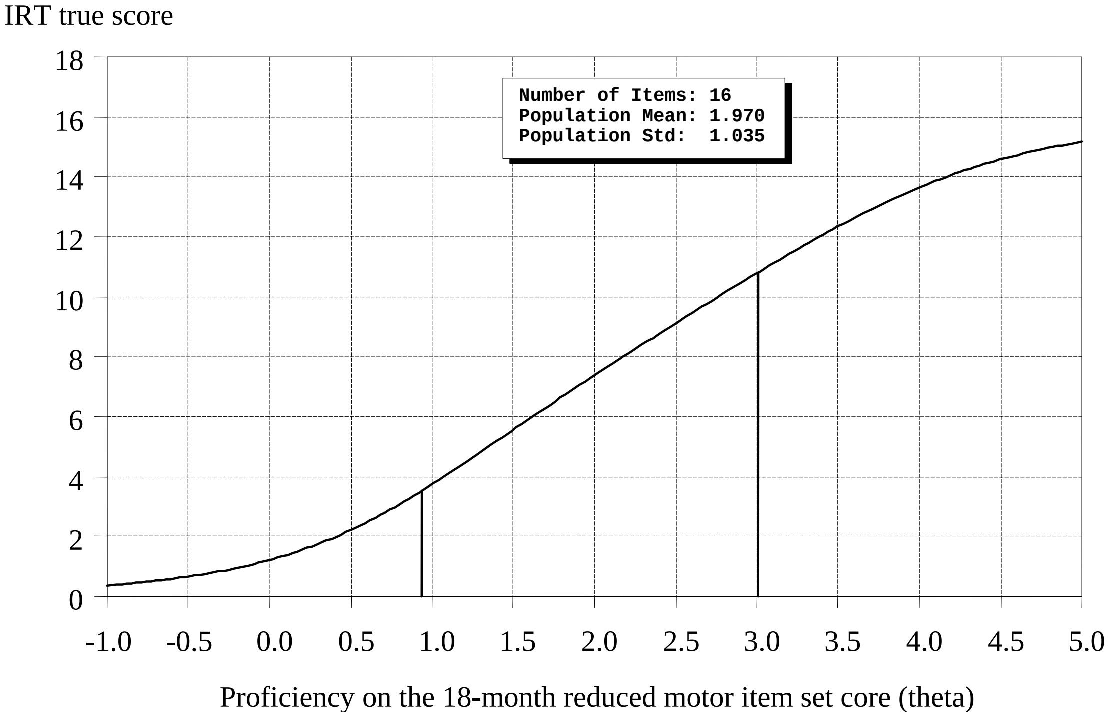
NOTE: The publisher does not provide data for all months of age in the calibration dataset but does provide recommended items for different months of age. The normal curve (dashed line) represents the projected latent distribution of the 26- to 28-month population and is included for illustration purposes.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

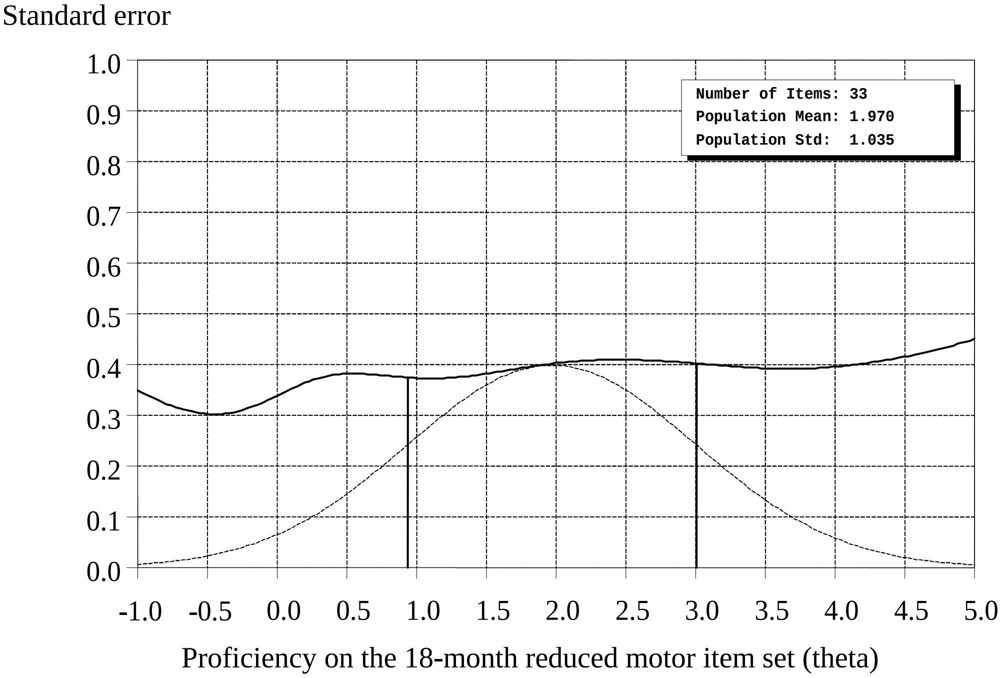
As with the mental scale, a child would never be administered all of these items. Neither would the basal or ceiling items be administered by themselves, but only after first administering the core item set. Consequently, a child can be administered either the core set of items alone, the core plus basal 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 percent x 17) + (16 percent x 24) + (16 percent x 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 items[9] (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 as 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.

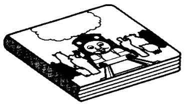
Where appropriate, boxes were also included that gave explicit warnings, such as “Don’t let the child put the beads in his mouth,” as well as troubleshooting instructions for problematic situations that can arise.

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.

Exhibit 2-4. Sample administration page, 18-month field test: 2003–04

## 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?<br>□ Check box if child said AT LEAST two 2-word sentences. | **Record what child said here:**<br>1. ______________________________<br>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.

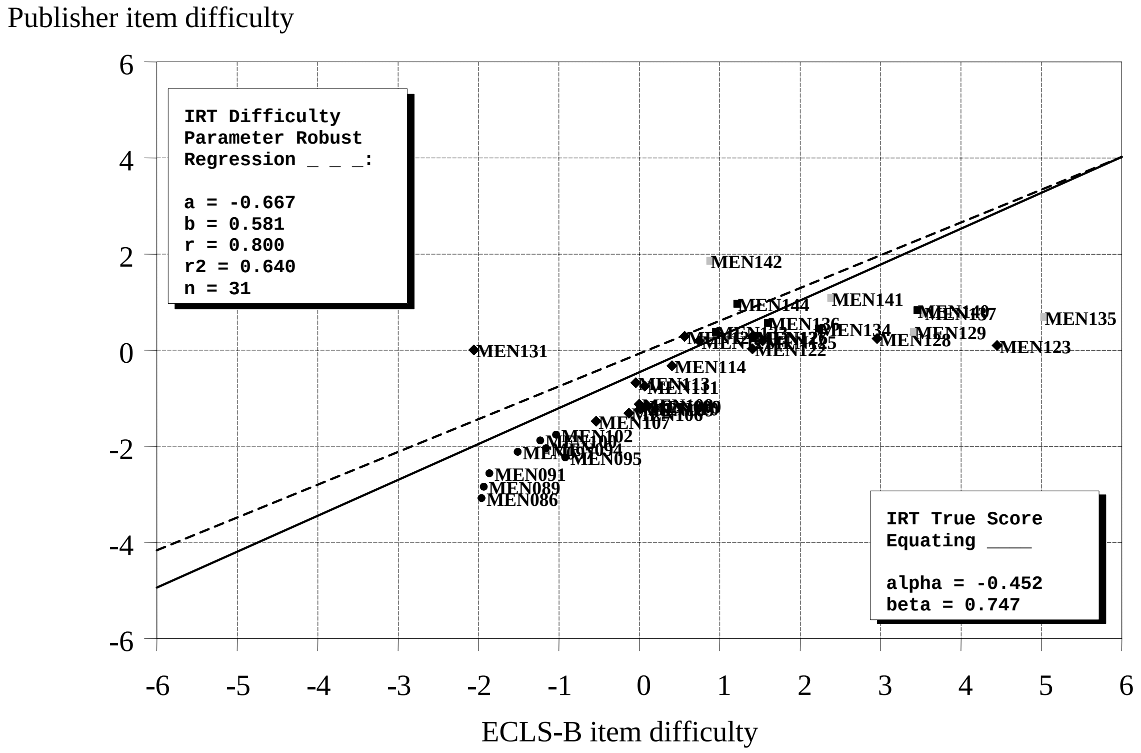
Of these excluded items, "Uses pads of fingertips to grasp pencil" was of the greatest concern. As a fine motor item, it should have been less sensitive to the skewed age distribution. It would appear from the data that interviewers were uncertain how to score this item and suggested that training on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

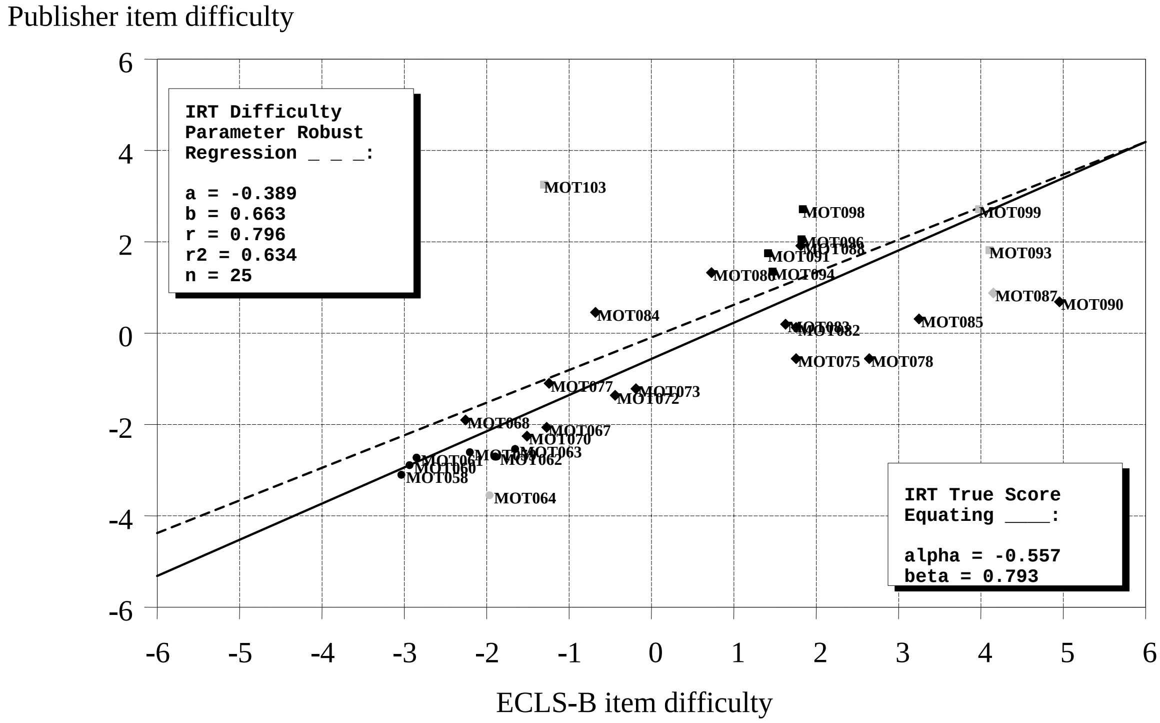


NOTE: a = regression intercept, when publisher item difficulty parameters are regressed on BSF-R item difficulty parameters; b = slope coefficient, when publisher item difficulty parameters are regressed on BSF-R item difficulty parameters; alpha = linear transformation of scale; beta = linear transformation 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



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.

# 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.

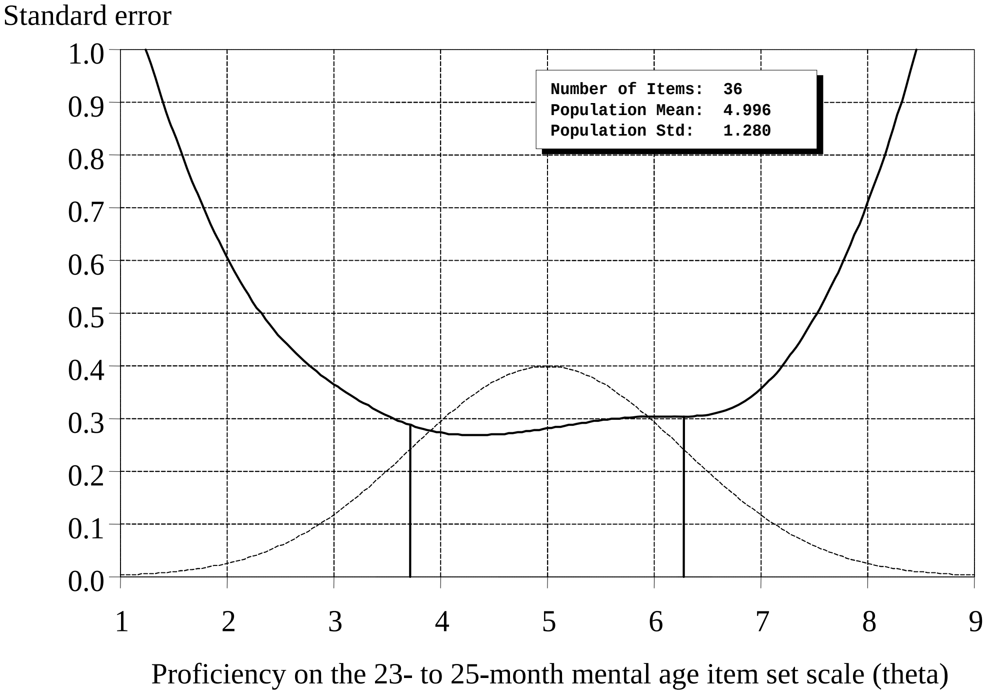
Once a list of candidate items that ranged from -3 to +3 standard deviations around the mean of the ability distribution for 2-year-olds was composed, items that were clearly not operationally feasible were eliminated. On the mental scale, for example, the item “Identifies objects in photographs” was 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.

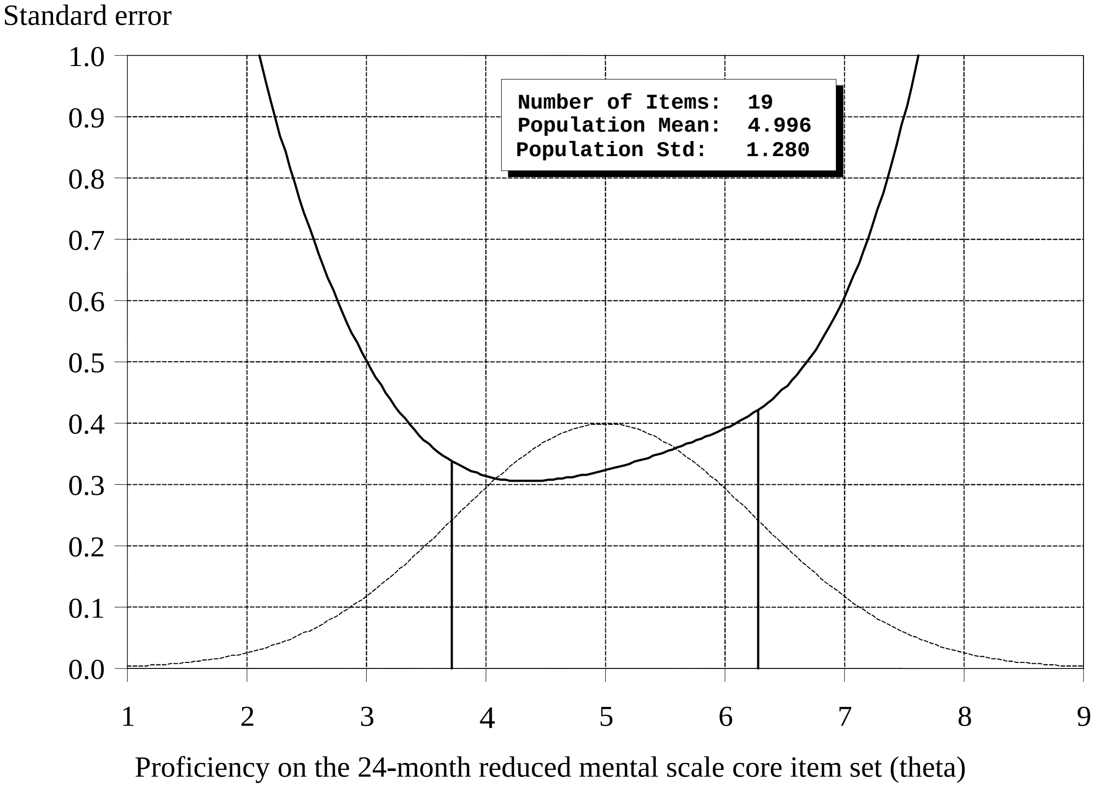
Figure 3-1. Standard error of measurement by proficiency level for the publisher-recommended 23- to 25-month age item set of the mental scale: Item Response Theory 2-parameter logistic item calibrations using publisher data: 1993



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.

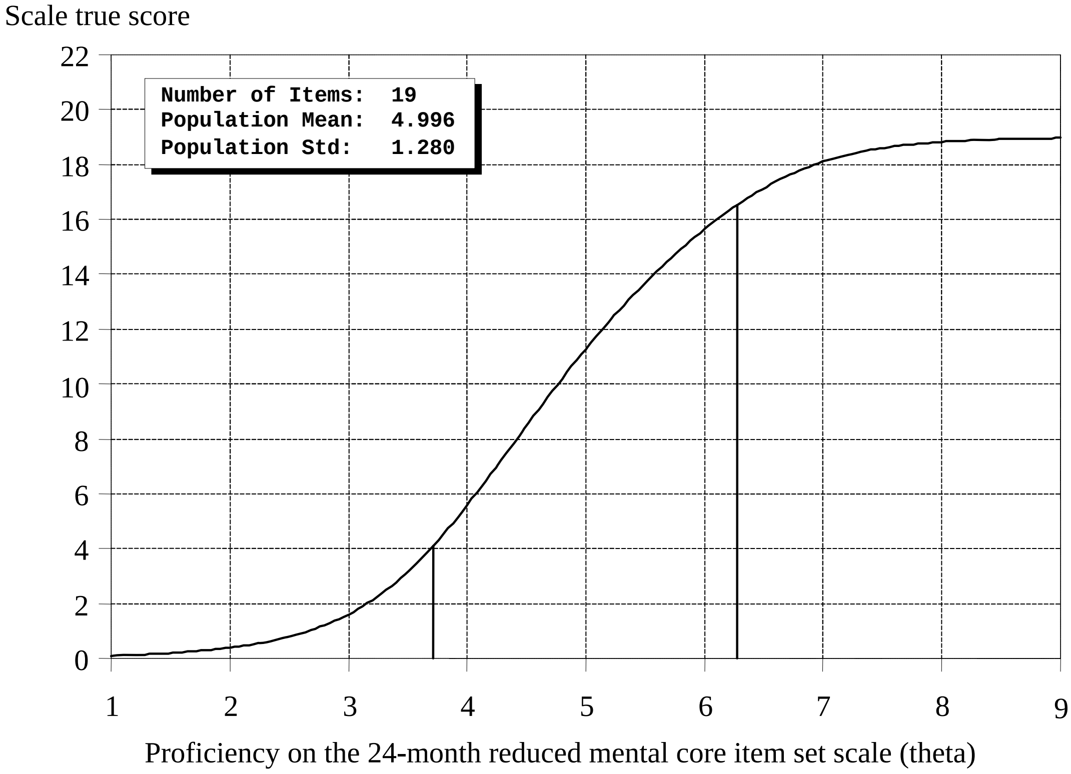
## 3.3 Creating the 2-year BSF-R Mental Scale Basal and Ceiling Sets

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.

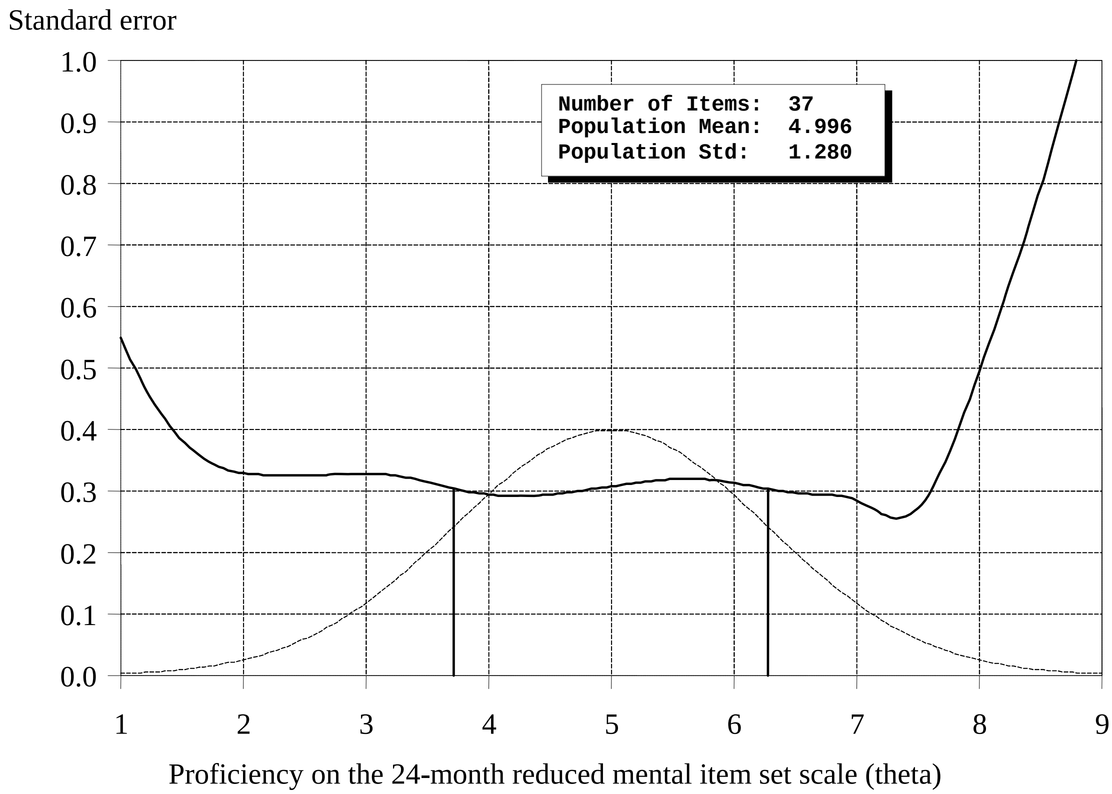
[4] Some items can be scored by observation during testing, but if the item is not observed, then it is administered.

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

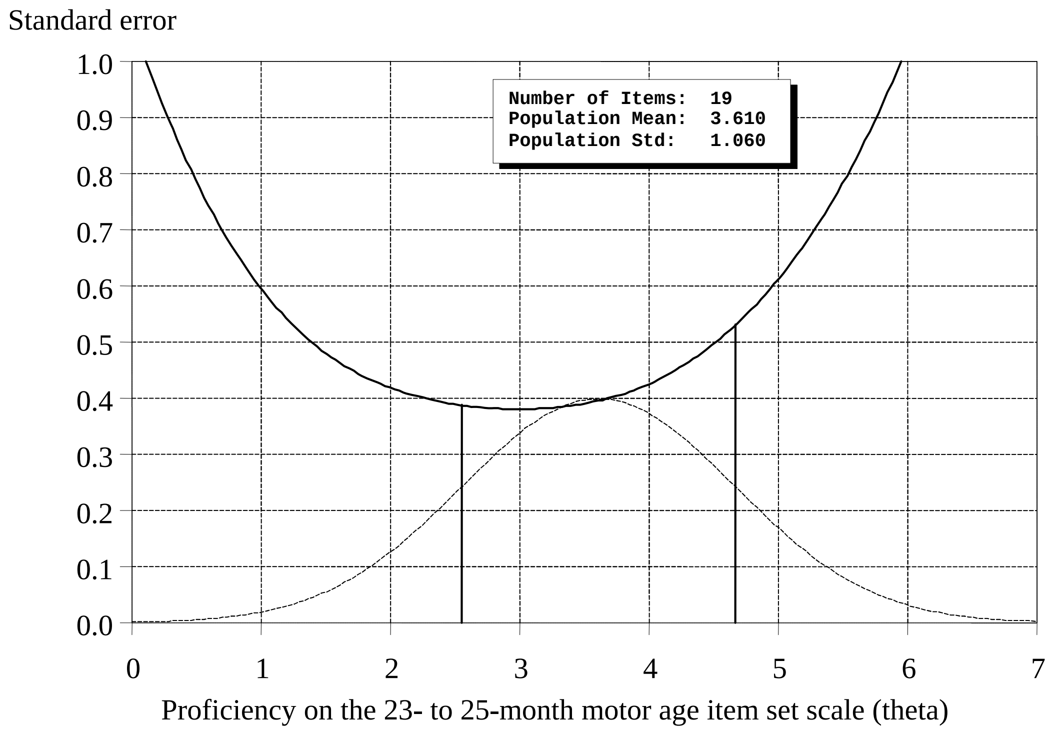


NOTE: Vertical lines descending from the solid curve to the x-axis indicate 1 standard deviation above the mean. Std = standard deviation. The normal curve (dashed line) represents the projected latent distribution of the 23- to 25-month population and is included for illustration purposes.
SOURCE: Publisher 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.

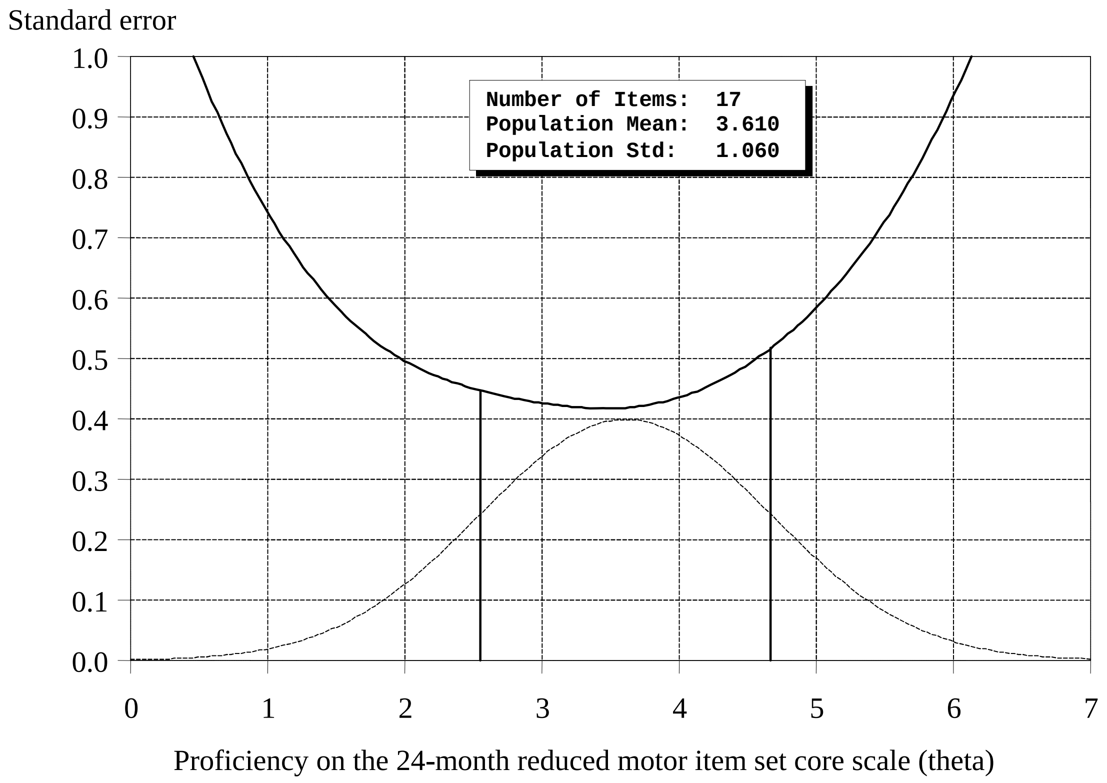
Figure 3-5. Standard error of measurement by proficiency level for the BSID-II motor scale 23- to 25-month age item set: IRT 2-parameter logistic item calibrations using publisher data: 1993



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.

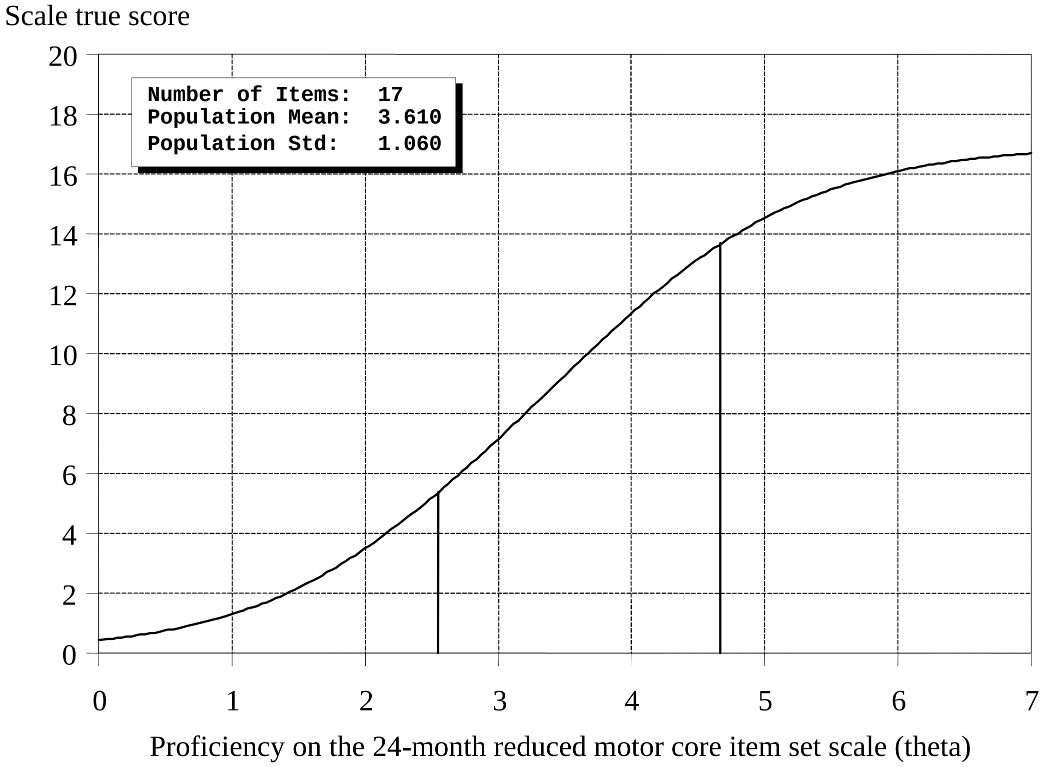
## 3.7 Creating the 2-Year BSF-R Motor Scale Basal and Ceiling Sets

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.

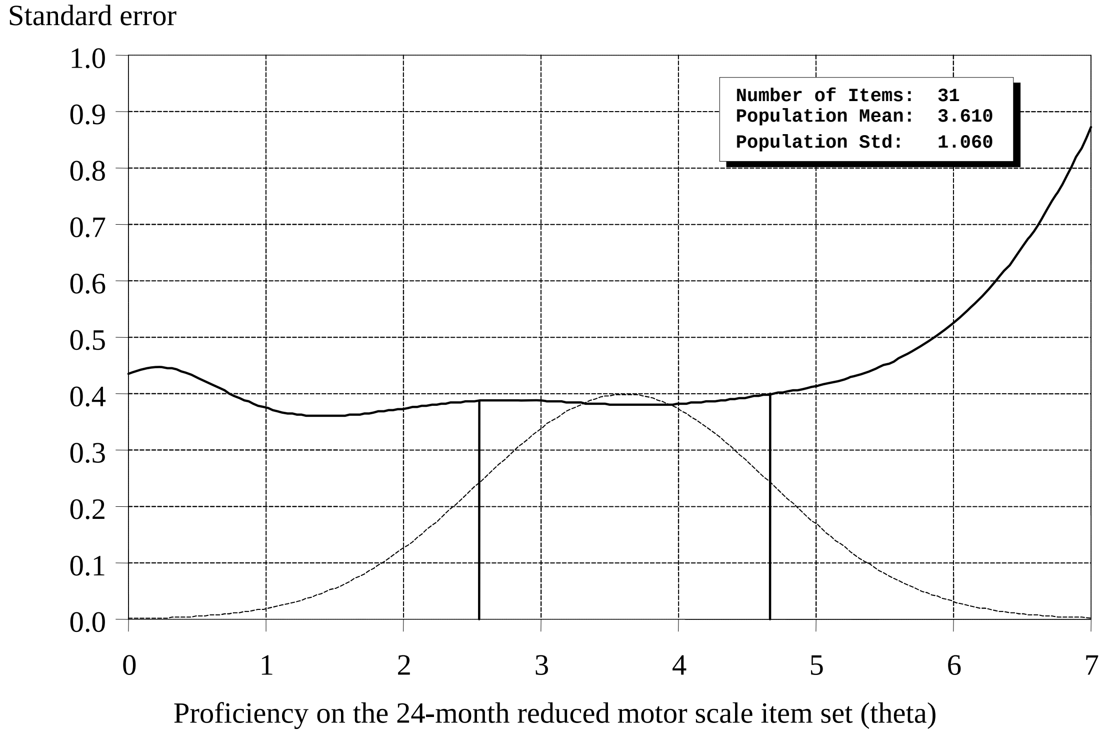
[3] Core score means that the score for this item is obtained during the administration of a core item.

[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.

# 4. CONCURRENT CALIBRATIONS AND EQUATING DESIGN FOR THE BSF-R IN THE 2-YEAR NATIONAL DATA COLLECTION

A shortened and streamlined version of the Bayley Scales of Infant Development, Second Edition (BSID-II) called the Bayley Short Form—Research Edition (BSF-R) was specially developed to assess child developmental status in the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B). One of the major justifications for using the Bayley scales in the ECLS-B was that they would produce results comparable with other child development studies that also report results for the BSID-II. Care was taken in selecting BSF-R item subsets so that ECLS-B results would be as consistent as possible with BSID-II. In particular, Mental Development Index (MDI) and Psychomotor Development Index (PDI) scores, developed by the publisher from a nationally representative standardization sample of children collected in 1991–92, could then also be reported in the ECLS-B.

For test construction purposes, the BSID-II standardization dataset was obtained from The Psychological Corporation, and all 178 mental and 111 motor items for infants between 1 and 42 months of age were calibrated using an Item Response Theory (IRT) two-parameter logistic (2-PL) model, a response model that was specifically chosen to highlight the more discriminating BSID-II items, which could then be identified for use in the ECLS-B. Although the initial item calibrations were undertaken by Westat, they are referred to here as *publisher calibrations* since these are based on the publisher's standardization dataset.

There are no weights on the standardization dataset. The standardization sample is representative of the national infant population and is considered to be self-weighting. The standardization dataset is comprised of the standardization sample and additional observations. None of these observations have case weights.

By contrast, the ECLS-B dataset is a stratified cluster sample based on unequal selection probabilities. Sample weights are used with the ECLS-B dataset so that it will then be representative of the national infant population in 2001. Thus, through calibration, scaling, scoring, and analysis and throughout this report, the standardization data were unweighted and the ECLS-B data were weighted.

## 4.1 Results of the BSF-R Adaptive Testing Strategy

Both the full BSID-II and BSF-R short forms were designed to be administered as adaptive tests. A core item set, appropriate for children in the target age group, was administered first. The raw score total for this core item set was then used to determine if additional basal or ceiling item sets should also have been administered. BSF-R adaptive tests followed procedures of administration similar to those used in BSID-II. The BSF-R diverged from the BSID-II primarily in its use of shortened core, basal, and ceiling item sets. The BSF-R was composed of shorter tests that were not designed to be strictly parallel tests.

Moreover, the BSF-R was specially adapted for home administration as part of household survey interviews conducted in the ECLS-B. The BSF-R was completed by interviewers guided by a standard schedule of task administrations, involving the structured presentation of stimulus material intended to elicit child responses. Still other items were scored based on observed child behavior occurring at any moment during assessment. Additionally, one or more items were scored from each task administration or observation. The first three sections of table 4-1 report the number of items in each of the basal, core, and ceiling item sets, followed by the number of task administrations and observations completed by interviewers before recording item responses. The sum of task administrations and observations does not equal the total number of items because in several instances more than one item was scored from a single task administration or observation. The difference between the sum of task administrations plus observations, and the total number of items, is shown in the last section of the table. This difference represents the number of items that cannot be considered entirely independent items.[1]

Although tests with a different number of items and other minor adaptations do not satisfy the rigorous requirements for test equating, tests based on the same item pool can often be calibrated on a common scale metric. Tests from the same item pool then yield unbiased ability estimates with the same central tendency but different standard errors. IRT procedures used in BSF-R design and development offered the prospect of producing comparable scores sharing the same scale metric used by the publisher. This metric was used to report BSF-R results, including model-based estimates of BSID-II raw scores and developmental index scores.

[1] Table 4-1 summarizes an exceedingly complex observational setting. This is because some item responses are recorded based on prior observation while requiring a separate administration on another occasion. It is even possible for a mental item to be scored from a motor administration. Except for item counts in the first section of the table, the numbers in other sections should be considered approximations rather than represent a full accounting of the situation encountered on each occasion.

Table 4-1. Number of BSF-R items, administrations, observations, and dependencies, by BSF-R scale, round of data collection, and item set: 2001–02 and 2003–04

| | 9-month data collection | | | 2-year data collection | | |
|---|---|---|---|---|---|---|
| Characteristic and item set | Total | Mental scale | Motor scale | Total | Mental scale | Motor scale |
| Number of items | | | | | | |
| Total | 66 | 31 | 35 | 69 | 37 | 32 |
| Basal | 21 | 9 | 12 | 18 | 9 | 9 |
| Core | 26 | 13 | 13 | 36 | 19 | 17 |
| Ceiling | 19 | 9 | 10 | 15 | 9 | 6 |
| Number of administrations | | | | | | |
| Total | 34 | 17 | 17 | 43 | 25 | 18 |
| Basal | 6 | 2 | 4 | 7 | 4 | 3 |
| Core | 18 | 9 | 9 | 26 | 14 | 12 |
| Ceiling | 10 | 6 | 4 | 10 | 7 | 3 |
| Number of observations | | | | | | |
| Total | 15 | 9 | 6 | 8 | 6 | 2 |
| Basal | 9 | 5 | 4 | 4 | 3 | 1 |
| Core | 2 | 2 | 0 | 3 | 2 | 1 |
| Ceiling | 4 | 2 | 2 | 1 | 1 | 0 |
| Number of dependencies | | | | | | |
| Total | 17 | 5 | 12 | 18 | 6 | 12 |
| Basal | 6 | 2 | 4 | 7 | 2 | 5 |
| Core | 6 | 2 | 4 | 7 | 3 | 4 |
| Ceiling | 5 | 1 | 4 | 4 | 1 | 3 |

NOTE: Item count: The number of items that are scored in each item set. Administrations: The number of task administrations. Each administration is defined as the structured presentation of the stimulus materials to obtain the child's response(s). Thus, multiple cores can be obtained from the same administration. Observations: Observations are items that do not require the structured presentation of stimulus materials but are scored by direct observation of the child's spontaneous behavior. Dependencies: A dependent item is redundant with another item and does not provide unique information about the child's ability and therefore does not increase construct representation. Item dependencies may also exacerbate any construct-irrelevant factors that may be associated with an item (e.g., prior familiarity with the item). The number of dependencies is calculated as the difference between the total item count and the sum of the number of administrations and observations.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

The BSF-R was designed so that most children received only the core item set, while other children received both the core item set plus either the basal or ceiling item set, depending on basal and ceiling decision rules based on the raw score total obtained with the core item set. Basal and ceiling decision rules were defined so that children within a standard deviation to either side of the age group mean ability estimate received only the core item set. In ideal circumstances, this would imply that about 68 percent of the children received only the core item set. Below a certain minimum core item set raw score value, another 16 percent were expected to receive the additional complement of basal items. Above a certain maximum core item set raw score value, another 16 percent were expected to receive the additional complement of ceiling items.

These percentages were expected to vary depending on the actual raw score values obtained with the core item sets. Basal and ceiling item sets contained 6 to 11 items that were specially selected to cover the child population well into the tails of each ability distribution. Table 4-2 shows how many children received the core, basal, and ceiling item sets on the BSF-R mental and motor tests. This shows that the item sets performed more or less as expected at 2 years. However, at 9 months, both the mental and motor distributions shifted upward toward the higher levels of ability, resulting in a greater use of both mental and motor ceiling items. While this resulted in a certain amount of inefficiency—in the sense that more items needed to be administered during field work—in principle, the ceiling item sets handled appropriately this need for a test with more difficult items.

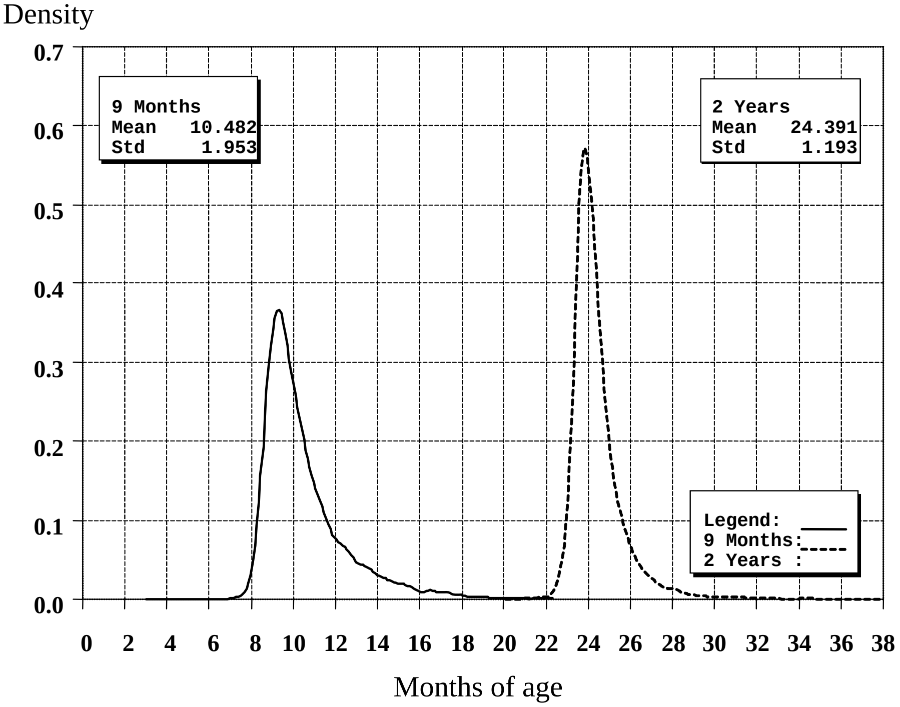
Table 4-2. Number and percentage of children, and mean ability estimates, by BSF-R scale, item set, and round of data collection: 2001–02 and 2003–04

| | 9-month data collection | | | 2-year data collection | | |
|---|---|---|---|---|---|---|
| BSF-R scale and item set | Number of children | Percent | Mean | Number of children | Percent | Mean |
| Mental scale | | | | | | |
| Total | 10,200 | 100.0 | -0.982 | 8,900 | 100.0 | 4.306 |
| Basal | 250 | 2.4 | -2.624 | 1,100 | 12.2 | 2.249 |
| Core | 5,650 | 55.4 | -1.466 | 6,150 | 69.1 | 4.221 |
| Ceiling | 4,300 | 42.2 | -0.982 | 1,650 | 18.7 | 5.967 |
| Motor scale | | | | | | |
| Total | 10,200 | 100.0 | -0.942 | 8,850 | 100.0 | 2.889 |
| Basal | 450 | 4.5 | -3.074 | 1,150 | 12.8 | 1.779 |
| Core | 6,300 | 62.0 | -1.592 | 7,150 | 80.7 | 2.966 |
| Ceiling | 3,400 | 33.4 | 0.553 | 600 | 6.5 | 4.114 |

NOTE: Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-1 shows why 9-month ability distributions shifted upward, resulting in the administration of fewer basal and more ceiling item sets than initially expected. This was because the age distribution at 9 months was shifted to the right and was highly skewed. Instead of being assessed at an average age of 9.5 months as initially expected, the first wave of infants were assessed at an average age of 10.5 months. For every additional month of age, mental ability estimates are expected to rise by fully 0.5 population standard deviations. In that case, only 7 percent required the basal and 31 percent required the ceiling items. However, it was not just that the age distribution was shifted to the right but also that it was highly skewed. In this case, even fewer infants were expected to receive the basal and even more received the ceiling item sets. Notice that for the second wave of assessments the age distribution was much closer to the expected 24.5 months of age, and the distribution was much less skewed. Thus, the age distribution of the ECLS-B sample at 9 months and 2 years was to a considerable extent responsible for the distribution of children who received the core, basal, and ceiling item sets.

Figure 4-1. Kernel density estimation for age distributions of children in the 9-month and 2-year ECLS-B data collections: 2001–02 and 2003–04



NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Because mental and motor development is explosive during infancy, infant age and development are closely related. This age-development relationship can be exploited during item calibration and scoring to improve the accuracy of item parameters and ability estimates. Observations were first clustered by age group, and the mean and standard deviation—representing the ability distribution in each age group—were used to condition group member ability estimates. The gains in precision obtained with multiple group IRT[2] are thought to be slight but help to ensure consistency when individual observations are scored.

Multiple group IRT (Bock and Zimowski 1997) was applied to ECLS-B item calibrations using Bilog-MG (Zimowski et al. 1997) and in-house software. The first set of software represents an industry standard and was useful for assessing the precision and accuracy of results. In-house software provided better graphics for visual inspection of item fit, together with almost unlimited flexibility during test equating and analysis. The two sets of software use multiple group IRT and produce results that are essentially identical. In multiple group IRT, item parameter values are estimated simultaneously together with the latent group ability distributions.

## 4.2 Examining the Potential of BSF-R Item Sets

An examination of the psychometric properties of BSF-R instruments began with an assessment of the potential of BSF-R item subsets before these were actually used in the ECLS-B. Among the 2,939 observations in the publisher's dataset between 1 and 42 months of age, 1,700 comprise the standardization sample, including a subset of 900 standardization observations between 8 and 30 months of age (see table 2-1). These age groups coincide most closely with the range of ages found in the ECLS-B. To examine the potential of BSF-R item sets, observations were compared after they were first scored with the full complement of BSID-II items and then again scored using BSF-R item subsets. While in principle this includes both the 9-month and 2-year item subsets, in practice each child was scored on the BSF-R items for which there were valid responses in the standardization dataset.[3]

[2] For further information about multigroup IRT, please refer to Bock, R.D., and Zimowski, M.F. (1997). Multiple Group IRT. In W.J. van der Linden and R.K. Hambleton, (Eds.), *Handbook of Modern Item Response Theory.* New York: Springer-Verlag.

[3] This implies that some children probably were scored on the 9-month item sets, others on the 2-year item sets, and still others on some items from both sets. As shown in table 4-8, the publisher standardization sample had insufficient number of observations to conduct separate analyses at 9 months and 2 years.

Having already calibrated all of the BSID-II items and scored all of the standardization sample observations with the full set of publisher item calibrations, the same observations were scored a second time using only the BSF-R item subsets. Publisher item calibrations were used on both occasions. Consequently, any differences encountered between the two sets of scores would have reflected a bias introduced by using the BSF-R item subsets. Results presented in table 4-3 permit a comparison of central tendencies, standard errors, and two measures of residual goodness of fit when the standardization sample was scored twice with different item subsets from publisher item calibrations.

When mean ability estimates obtained with the full BSID-II were compared with means obtained using BSF-R item subsets, these were expected to yield unbiased estimates of average ability. Mean expected a posteriori (EAP) ability estimates reported in table 4-3 were virtually identical on the mental scale and within a 10th of a population standard deviation on the motor scale (see section 5.2 for a more detailed description of the EAP). This was expected since the same standardization sample observations and publisher item calibrations were used to obtain both sets of means. Nevertheless, the results supported expectations. The central tendencies obtained with BSF-R item subsets faithfully reproduced those obtained using the full complement of BSID-II items.

Table 4-3. Descriptive statistics for 900 standardization sample observations scored with publisher calibrations using both the full BSID-II and BSF-R item sets: 2001–02 and 2003–04

| Scale | Full BSID-II | BSF-R item subset[1] |
|---|---|---|
| Mental | | |
| Mean EAP ability estimate | 2.511 | 2.512 |
| Mean EAP standard error | 0.282 | 0.345 |
| Information-weighted mean square residual goodness of fit—Infit | 0.954 | 0.915 |
| Outlier-sensitive mean square residual goodness of fit—Outfit | 0.964 | 0.931 |
| Motor | | |
| Mean EAP ability estimate | 1.648 | 1.634 |
| Mean EAP standard error | 0.349 | 0.399 |
| Information-weighted mean square residual goodness of fit—Infit | 0.935 | 0.891 |
| Outlier-sensitive mean square residual goodness of fit—Outfit | 0.884 | 0.835 |

[1] Includes both 9-month and 2-year item subsets.

NOTE: Publisher standardization dataset observations 8 through 30 months of age that corresponding most closely with age groups found in the ECLS-B sample. EAP = expected a posteriori.

SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

For both the mental and motor scales, BSF-R instruments yielded a somewhat larger standard error when compared with the corresponding standard errors obtained from scores using BSID-II item sets. The average standard error for the 900 standardization sample observations on the BSF-R mental subset was 0.345 population standard deviation, which compared with 0.282 for the full BSID-II mental set. The average standard error on the BSF-R motor item set was 0.399, which compared with 0.349 for the full BSID-II motor set. Indeed, the somewhat larger standard errors obtained with the BSF-R instruments were expected since these were obtained using smaller item subsets. Although no single child would ever be administered all 178 mental items or all 111 motor items, BSID-II core, basal, and ceiling sets were invariably larger than those included in BSF-R instruments.

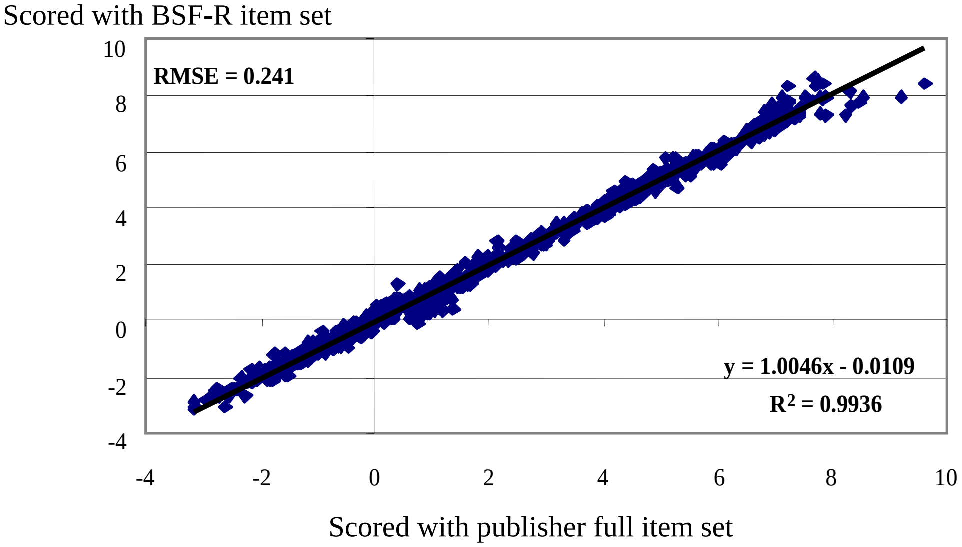
These comparisons were reassuring in the sense that somewhat larger standard errors for the BSF-R short forms had been expected. The BSF-R tests were never intended to be strictly parallel tests yielding virtually identical ability estimates and similar standard errors found in BSID-II. Instead, using publisher item parameters, the BSF-R item subsets were expected to perform like τ-equivalent tests, producing essentially identical ability estimates but somewhat larger standard errors.[4]

This same analysis can be repeated across the entire ability range for standardization sample observations between the ages of 8 and 30 months. Linear relationships between observations scored with the full BSID-II and scored again with the shorter BSF-R item subsets are shown in figures 4-2 and 4-3. To the extent that it was possible for BSF-R item subsets to produce results that were identical to those produced with the full BSID-II, the EAP ability estimates would have aligned themselves precisely along a straight line having an origin of zero and slope of unity.

Indeed, the two figures show that the central tendency of the relationship between the two sets of ability estimates had an origin very close to zero and a slope very close to unity. The $r^2$ coefficients reported in the figures are also close to unity, suggesting that the relationship between the two sets of scores was nearly perfect. However, the ability range between 8 and 30 months of age is so large that the $r^2$ statistics may be somewhat misleading. A better measure of the imperfection in measurement is provided by the root mean squared error (RMSE) reported in each of the figures. These values show that the expected error of estimation obtained with the reduced item subsets was approximately one-quarter of a population standard deviation (RMSE = 0.241 for the BSF-R mental and RMSE = 0.224 for the motor). The average error is exceedingly small, suggesting that, under clinical conditions, the BSF-R item subsets were capable of predicting BSID-II ability estimates with considerable precision across a broad range of ability.

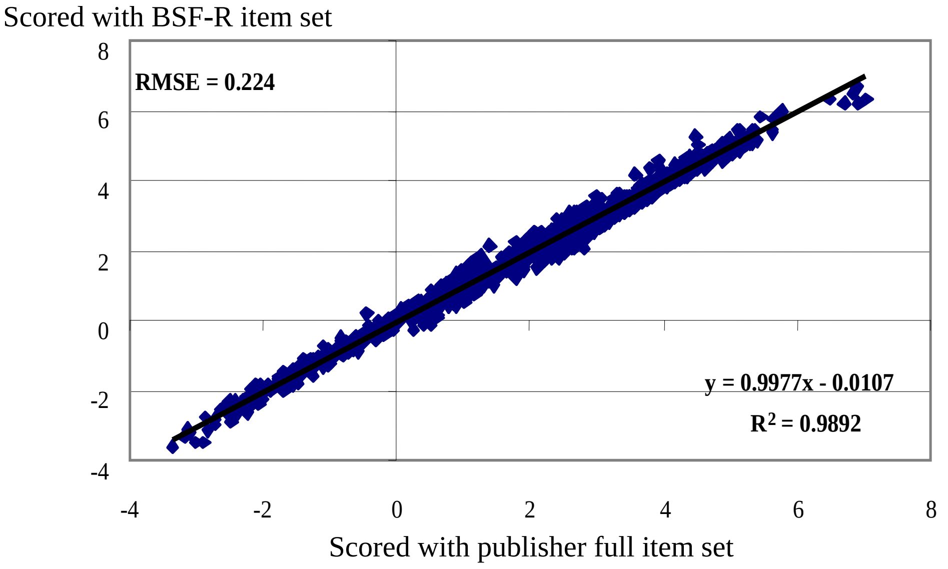
[4] The *τ*-equivalent refers to measurements that have the same true scores but possibly different standard errors (Lord and Novick 1968).

Figure 4-2. Expected a posteriori ability estimates, using fixed publisher item calibrations throughout; standardization sample observations scored first with the full BSID-II mental (*x* axis) and then with the BSF-R mental item set (*y* axis): 1993



NOTE: RMSE = root mean squared error; $R^2$ = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Figure 4-3. Expected a posteriori ability estimates, using fixed publisher item calibrations throughout; standardization sample observations scored first with the full BSID-II motor (*x* axis) and then with the BSF-R motor item set (*y* axis): 1993



NOTE: RMSE = root mean squared error; $R^2$ = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Measurement accuracy was also assessed with person-fit analyses of individual response vectors. Person-fit indices showed to what extent a response pattern was considered typical. One would ordinarily expect an examinee to obtain correct responses to easy items, provide correct and incorrect responses to items in the vicinity of his or her ability, and obtain incorrect responses to hard items. Fit statistics measure the extent to which a response pattern contains surprisingly correct responses to difficult items or surprisingly incorrect responses to easy items. As such, fit statistics show the extent to which the data are found to be appropriate for the IRT model. To assess this issue, average person fit statistics are also reported in table 4-3.

Both outlier sensitive (Outfit) and information-weighted (Infit) mean square statistics are reported in table 4-3. Outfit is based on the sum of squared residuals normalized by the variance around its expectation. The disadvantage of this statistic is that it is quite sensitive to unexpected responses to items that are much too easy or much too difficult. Infit is an information-weighted measure that gives less weight to remote items in determining on the magnitude of the fit statistic (Linacre and Wright 1994). The expected value for the mean square residual on both of these indices is 1.0. Departures from expectation are represented by values noticeably above or below unity, where large values represent excessive noise and small values represent insufficient stochastic variation needed for useful measurement. For reasonably large samples, fit statistics greater than 1.1 indicate departures from expected response patterns that require further attention (Smith, Schumacker, and Bush 1998).

Applying these same criteria, all of the Infit statistics reported in table 4-3 are slightly less than unity. An Infit index of 0.9 implies that there is 10 percent less randomness than expected among item responses that closely matched the respondent’s ability level. In this case, individual responses were too predictable, fit the response model too closely, and provided redundant information when these observations were scored with IRT. In the case of the Bayley, this may have resulted from coding several item responses from a single task administration. The assessor effectively behaved as if he or she were imputing responses rather than recording behavior observed after independent trials. Outfit statistics for both motor item sets were satisfactory but again showed evidence of redundant information contained in the item responses.[5]

In general, Infit values in excess of the criterion value of 1.1 are a more serious problem than Outfit values in excess of the same criterion value. This is because high Infit values show that the data fail

[5] Although it can be safely assumed that standardization sample observations followed publisher recommendations in applying basal and ceiling item sets, no effort was made here to assure that responses to BSF-R items follow basal and ceiling rules prescribed for BSF-R administration.

to fit the response model at the point where they are most needed to estimate a person's level of ability. High Outfit values are easier to manage because in a worst case scenario suspect responses could be replaced with missing values without serious impact on ability estimates.

Table 4-4. Number and percentage distribution of 900 standardization sample observations scored with publisher calibrations using both the full BSID-II and combined 9-month and 2-year BSF-R item sets, by level of outfit and scale: 1993

| | | Full BSID-II | | BSF-R subsets[1] | |
|---|---|---|---|---|---|
| Fit | Level of outfit | Number | Percent | Number | Percent |
| Mental scale | | | | | |
| Total | $0 \le y < \infty$ | 900 | 100.0 | 900 | 100.0 |
| Excellent | $0 \le y < 1$ | 618 | 68.7 | 663 | 73.7 |
| Acceptable | $1 \le y < 3$ | 263 | 29.2 | 200 | 22.2 |
| Problematic | $3 \le y < 5$ | 19 | 2.1 | 29 | 3.2 |
| Unacceptable | $5 \le y < \infty$ | 0 | # | 8 | 0.9 |
| Motor scale | | | | | |
| Total | $0 \le y < \infty$ | 900 | 100.0 | 900 | 100.0 |
| Excellent | $0 \le y < 1$ | 655 | 72.8 | 668 | 74.2 |
| Acceptable | $1 \le y < 3$ | 236 | 26.2 | 225 | 25.0 |
| Problematic | $3 \le y < 5$ | 8 | 0.9 | 4 | 0.4 |
| Unacceptable | $5 \le y < \infty$ | 1 | 0.1 | 3 | 0.3 |

# Rounds to zero.

[1] Includes both 9-month and 2-year item subsets. Outfit = outlier-sensitive mean squared residual goodness of fit.

NOTE: Standardization dataset observations 8 through 30 months of age. Detail may not sum to total because of rounding.

SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Fit statistics are generated for each observation, which implies that observations should be examined individually. Table 4-4 reports sample frequencies for the Outfit statistic broken down into four categories. Outfit < 1 represents model fit that is considered extremely good but may suggest redundant information. This is followed by a category where 1 < Outfit < 3, which may show some evidence of misfit but is usually considered satisfactory. Outside this range, 3 < Outfit < 5 would be considered problematic and elicit attention to individual cases. At the extreme, where Outfit > 5, model fit is considered unacceptable, and individual cases should probably not be given score values. The table shows that the 900 observations between 8 and 30 months of age were generally well represented by publisher item calibrations, with very few observations that might be considered problematic. The full complement of BSID-II standardization sample observations essentially escape being labeled unacceptable, with the possible exception of a single observation on the motor scale. Generally speaking, subsets of BSF-R items performed almost as well but revealed a handful of observations on both the mental and motor scales that probably should not have been given score values.

## 4.3 BSF-R Compatibility

The performance of BSF-R instruments used in the ECLS-B fieldwork were then considered. Assuming that IRT parameter invariance properties hold, then it would have been possible to score ECLS-B observations directly using publisher item calibrations.[6] This possibility was examined using ECLS-B longitudinal data collected at 9 months and 2 years of age. As reported in table 4-5, the ECLS-B sample consisted of approximately 10,215 children assessed at two points in time. This includes 10,197 children assessed on the mental scale and 10,163 on the motor scale during the 9-month data collection and 8,912 assessed on the mental scale and 8,824 assessed on the motor scale during the 2-year data collection. The difference between the total assessed at 9 months and 2 years is largely a reflection of the 1,359 children who were not assessed at the 2-year data collection, since only 8 children who had not been assessed at 9 months were assessed at 2 years. Frequency counts for completed assessments showed that nearly all who took the mental assessment also took the motor assessment. Due to delays in scheduling fieldwork, comparatively few children were assessed prior to 9 months or prior to 2 years of age, whereas many more were assessed as many as several months beyond the expected ages.

Given that the full set of BSID-II items had already been calibrated using the publisher standardization dataset, the possibility of using publisher item parameters to score ECLS-B observations was logically considered. Assuming that the ECLS-B data were to fit the publisher IRT model, all of the resulting EAP ability estimates could then be easily reported on the scale metric used by the publisher. This would obviate any need for an independent set of ECLS-B item calibrations or any kind of scale equating. Instead, BSF-R IRT ability estimates obtained with publisher calibrations could be used to calculate publisher IRT true scores, each of which would provide a model-based estimate of the BSID-II number-right raw score. Model-based estimates of raw scores or developmental index scores reported in the ECLS-B would then be directly compared with BSID-II results reported elsewhere.

[6] Item difficulty parameter estimates from separate calibrations align themselves along a straight line, indicating that a simple transformation of origin and scale is all that is needed to place one set of items on the same scale metric as the other item set. This constitutes a rigorous test of IRT parameter invariance properties.

Table 4-5. Cross-classification of number of children assessed/not assessed at 9-month and 2-year round of data collection, by BSF-R scale: 2001–02 and 2003–04

| | 2 years | | |
|---|---|---|---|
| 9 months | Not assessed[1] | Assessed | Total |
| Mental scale | | | |
| Total | 1,300 | 8,900 | 10,200 |
| Not assessed | # | # | # |
| Assessed | 1,300 | 8,900 | 10,200 |
| Motor scale | | | |
| Total | 1,350 | 8,850 | 10,200 |
| Not assessed | # | 50 | 50 |
| Assessed | 1,350 | 8,800 | 10,150 |

# Rounds to zero.

[1] Not assessed includes children who failed to complete two-thirds of the core item set and by decision of NCES were not scored. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

For this to work properly, it was first necessary to demonstrate that ECLS-B data conformed to the publisher IRT model. Implicitly, this would test the hypothesis that ECLS-B instruments, interviewers, and administration procedures, when used in a household interview setting, resulted in recorded responses that were consistent with those obtained with the full BSID-II when this was used in a clinical setting. Evidence supporting this hypothesis would provide a strong argument to support the validity of ECLS-B measures since these could then be shown to produce $\tau$-equivalent results essentially identical to those obtained under clinical conditions. If the evidence *failed* to support this hypothesis, this finding would imply that ECLS-B instruments, interviewers, and procedures produced results that were inconsistent with those of the full BSID-II when used under clinical conditions. However, in this eventuality, it would still be possible to calibrate BSF-R items on a common-scale metric consistent with the BSID-II.

For this first experiment, only ECLS-B data collected using BSF-R instruments, interviewers, and administration procedures were used. ECLS-B data were scored directly using publisher item calibrations. Experiment results are reported in table 4-6. These results show that ECLS-B mental item responses are fairly inconsistent in the vicinity of the child’s ability and very inconsistent on items far removed from the child’s ability. With mean Infit indices ranging from 1.364 at 9 months to 1.513 at 2 years, this implies that, *from the BSID-II perspective*, there was considerable noise in the vicinity of child ability. Interpreted literally, this index shows that there is anywhere from 36 percent to 50 percent more random noise in ECLS-B data than would be expected had the data actually conformed to publisher item calibrations. This finding was of critical importance because it showed that ECLS-B data failed to fit the response model at the point where they were most needed to estimate child ability.

Table 4-6. Mean fit indices for ECLS-B observations scored directly with publisher calibrations, by BSF-R scale and round of data collection: 2001–02 and 2003–04

| | | Mean fit values | |
|---|---|---|---|
| ECLS-B subsample | Mean Squared Residual Fit Index | BSF-R mental item subset | BSF-R motor item subset |
| 9 months | Information-weighted mean squared residual goodness of fit—Infit | 1.364 | 1.096 |
| | Outlier-sensitive mean squared residual goodness of fit—Outfit | 2.045 | 1.267 |
| 2 years | Information-weighted mean squared residual goodness of fit—Infit | 1.513 | 1.081 |
| | Outlier-sensitive mean squared residual goodness of fit—Outfit | 2.204 | 1.125 |

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

There seems to be an even larger problem with items that were either too easy or too difficult for the child, resulting in misjudgments when these item responses were recorded. Here, Outfit values slightly in excess of 2.0 that might not cause great concern with individual observations were considered unacceptably large for a group mean.

Both Infit and Outfit indices were more satisfactory in the case of the motor scale, however, the Infit means were close to the critical value of 1.1. Outfit indices for the motor scale were somewhat larger. Collectively, these indices showed that ECLS-B data do not fit the publisher response models as well as one would have liked. These findings suggest that the ECLS-B data required their own set of item calibrations and an appropriate equating design so that test results could be reported on the publisher scale metric.

Outfit frequencies for the ECLS-B sample shown in table 4-7 confirm what has already been stated regarding mean fit indices. Although model fit on the mental test was either excellent or acceptable for the majority of observations when these were scored using publisher calibrations, model fit was either problematic or unacceptable for large numbers of other observations. Motor fit was generally satisfactory at 2 years but not entirely satisfactory for an appreciable number of observations at 9 months.

Table 4-7. Number and percentage distribution of ECLS-B sample observations scored with publisher calibrations, using combined 9-month and 2-year BSF-R item sets, by level of outfit and scale: 2001–02 and 2003–04

| | | BSF-R mental scale | | BSF-R motor scale | |
|---|---|---|---|---|---|
| Fit | Level of outfit | Number | Percent | Number | Percent |
| 9 months | | | | | |
| Total | $0 \le y < \infty$ | 10,200 | 100.0 | 10,200 | 100.0 |
| Excellent | $0 \le y < 1$ | 4,100 | 39.9 | 5,950 | 58.3 |
| Acceptable | $1 \le y < 3$ | 4,050 | 39.9 | 3,350 | 33.1 |
| Problematic | $3 \le y < 5$ | 1,200 | 11.9 | 600 | 6.1 |
| Unacceptable | $5 \le y < \infty$ | 850 | 8.3 | 250 | 2.6 |
| 2 years | | | | | |
| Total | $0 \le y < \infty$ | 8,950 | 100.0 | 8,900 | 100.0 |
| Excellent | $0 \le y < 1$ | 1,600 | 17.7 | 4,050 | 45.7 |
| Acceptable | $1 \le y < 3$ | 5,200 | 58.1 | 4,750 | 53.4 |
| Problematic | $3 \le y < 5$ | 1,500 | 16.6 | 50 | 0.8 |
| Unacceptable | $5 \le y < \infty$ | 700 | 7.7 | # | 0.1 |

# Rounds to zero.

NOTE: Outfit = outlier-sensitive mean squared residual goodness of fit. Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

These findings are understandable given the unique set of circumstances encountered in ECLS-B fieldwork. It was reasonable to expect that a household interview setting would be different from a clinical setting. In a household interview, space sufficient for an assessment is often lacking, lighting is frequently inadequate for careful observation, and there is little opportunity to regulate the conditions in which to conduct an assessment. Nor was it expected that laymen interviewers with limited experience in child development would be able to replicate clinical measures in the absence of the specialized knowledge and expertise expected of trained clinicians. Although it had already been demonstrated that BSF-R item subsets were capable of providing unbiased estimates of measurement outcomes obtained with the full BSID-II, it was expected that the ECLS-B experience would yield somewhat different results due to fieldwork conditions and adaptations introduced to simplify administration of some of the items. However, even in these circumstances, it was possible to calibrate ECLS-B item subsets on the publisher scale metric.

## 4.4 BSF-R Conditioning

The task of calibrating the BSF-R item subsets on a common scale metric using ECLS-B data was then considered. Table 4-8 reports frequency counts for both the ECLS-B sample and publisher standardization dataset, broken down by months of age. The ECLS-B sample contained large numbers of assessments at two points in time. This age breakdown highlights ECLS-B observations intended for assessments at 9 months and 2 years of age. Due to the usual complexity of scheduling interviews, the age distribution on both occasions was highly skewed. Although some children in the sample were easily located and promptly interviewed, many others could only be interviewed after a series of scheduling delays. Thus, the age distribution at 9 months became skewed and waned at about the same age where 2-year assessments began. The challenge was to find a satisfactory means of placing scores for all these children on a common scale metric. A consistent scale metric is required for the longitudinal analysis of ECLS-B data.

With age distributions such as these, there was little opportunity to use item linkages between the 9-month and 2-year BSF-R tests to establish a common vertical scale. In fact, there were only two common items linking the two BSF-R mental tests and only a slightly more expressive number of eight item linkages between the two motor scales. In any case, these were fairly atypical items that served as ceiling items at 9 months and basal items at 2 years. With relatively few items and smaller numbers of respondents for these items, there was little opportunity to develop an equating design based on common item linkages between BSF-R tests at 9 months and 2 years.

By contrast, the strength of the publisher dataset lay not so much in the number of standardization sample and other observations in this dataset, but rather with the strategic positioning of these observations over such an extensive range of infant ability found between 1 and 42 months of age.[7] This design assures the largest possible number of observations linking adjacent age item sets. In fact, there is an average of 633 (± 235 observations) for each mental item, ranging from a minimum of 257 to a maximum of 1,130 observations used to calibrate the publisher mental scale. There is an average of 564 (± 228 observations) for each motor item, ranging from a minimum of 174 to a maximum of 1,031 observations used to calibrate the publisher motor scale. For both the mental and motor scale, these observations provide a solid string of items calibrated across the widest possible range of infant development, assured by the extraordinary age variation found between 1 and 42 months of age.

[7] The publisher's standardization sample contained 100 observations for each of 17 selected age groups. The 1,700 standardization sample observations are complemented by an additional 1,239 observations of other infants. The higher percentage of basal items administered to this second group suggests that perhaps 4.5 percent of these observations show some evidence of deficient ability. The standardization sample and other observations in the combined sample of 2,939 observations were used to calibrate publisher item sets, affording the largest possible number of item responses linking adjacent age item sets.

Table 4-8. Frequency count for ECLS-B longitudinal sample and publisher standardization dataset, by test and months of age: 2001–02 and 2003–04

| Months | Mental scale | | | | | Motor scale | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| of | ECLS-B | | Publisher | | | ECLS-B | | Publisher | | |
| age | 9-month | 2-years | Stdz | Other | Total | 9-month | 2-years | Stdz | Other | Total |
| Total | 10,200 | 8,900 | 1,700 | 1,250 | 22,050 | 10,150 | 8,850 | 1,700 | 1,250 | 21,950 |
| 1 | † | † | 100 | 50 | 150 | † | † | 100 | 50 | 150 |
| 2 | † | † | 100 | 50 | 150 | † | † | 100 | 50 | 150 |
| 3 | † | † | 100 | 50 | 150 | † | † | 100 | 50 | 150 |
| 4 | # | † | 100 | # | 150 | # | † | 100 | # | 150 |
| 5 | # | † | 100 | # | 100 | # | † | 100 | # | 100 |
| 6 | 50 | † | 100 | # | 150 | 50 | † | 100 | # | 150 |
| 7 | 150 | † | † | † | 150 | 150 | † | † | † | 150 |
| 8 | 850 | † | 100 | 50 | 950 | 850 | † | 100 | 50 | 950 |
| 9 | 3,400 | † | † | † | 3,400 | 3,350 | † | † | † | 3,350 |
| 10 | 2,650 | † | 100 | # | 2,800 | 2,650 | † | 100 | # | 2,800 |
| 11 | 1,300 | † | † | † | 1,300 | 1,300 | † | † | † | 1,300 |
| 12 | 700 | † | 100 | 150 | 950 | 700 | † | 100 | 150 | 950 |
| 13 | 450 | † | † | † | 450 | 450 | † | † | † | 450 |
| 14 | 250 | † | † | † | 250 | 250 | † | † | † | 250 |
| 15 | 150 | 1 | 100 | 100 | 350 | 150 | # | 100 | 100 | 350 |
| 16 | 100 | † | † | † | 100 | 100 | † | † | † | 100 |
| 17 | 100 | † | † | † | 100 | 100 | † | † | † | 100 |
| 18 | 50 | # | 100 | 50 | 200 | 50 | # | 100 | 50 | 200 |
| 19 | # | # | † | † | # | # | # | † | † | # |
| 20 | # | 50 | † | † | 50 | # | 50 | † | † | 50 |
| 21 | # | 250 | 100 | 100 | 450 | # | 250 | 100 | 100 | 450 |
| 22 | # | 450 | † | † | 450 | # | 400 | † | † | 450 |
| 23 | † | 1,450 | † | † | 1,450 | † | 1,450 | † | † | 1,450 |
| 24 | † | 4,150 | 100 | 200 | 4,500 | † | 4,150 | 100 | 200 | 4,450 |
| 25 | † | 1,600 | † | † | 1,600 | † | 1,600 | † | † | 1,600 |
| 26 | † | 550 | † | † | 550 | † | 550 | † | † | 550 |
| 27 | † | 200 | 100 | 50 | 350 | † | 200 | 100 | 50 | 350 |
| 28 | † | 100 | † | † | 100 | † | 100 | † | † | 100 |
| 29 | † | 50 | † | † | 50 | † | 50 | † | † | 50 |
| 30 | † | # | 100 | 100 | 250 | † | # | 100 | 100 | 250 |
| 31 | † | # | † | † | # | † | # | † | † | # |
| 32 | † | # | † | † | # | † | # | † | † | # |
| 33 | † | # | † | † | # | † | # | † | † | # |
| 34 | † | # | † | † | # | † | # | † | † | # |
| 36 | † | # | 100 | 150 | 250 | † | # | 100 | 150 | 250 |
| 37 | † | † | † | † | † | † | # | † | † | # |
| 38 | † | # | † | † | # | † | # | † | † | # |
| 42 | † | † | 100 | 50 | 150 | † | † | 100 | 50 | 150 |

† Not applicable.

# Rounds to zero.

NOTE: Stdz: Publisher standardization sample. Other: Nonstandardization sample observations included in the publisher dataset. BSID ages for ECLS-B observations rounded to nearest whole number. Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04; Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Given the ECLS-B design and available publisher data, the best approach to build a comparable vertical scale metric to link the 9-month and 2-year samples was to use the publisher standardization dataset as a bridge linking the two data collections. Frequency counts for the publisher standardization sample are also reported in table 4-8. The standardization sample had fewer observations than the ECLS-B sample, but these were strategically positioned at regular intervals all the way from 1 to 42 months of age. Collectively, across all age groups, the standardization sample contained information on the full complement of 178 mental and 111 motor items. By contrast, designed for use at 9 months and 2 years, BSF-R instruments were based on item subsets that include 66 mental and 59 motor items. Each of these items provided a linkage between BSF-R short forms and the full BSID-II. These linkages were used to establish a consistent scale metric between BSF-R tests at 9 months and 2 years of age and to establish an ECLS-B scale metric that was consistent with publisher documentation.

Scale equating in the ECLS-B was approached in the context of a Non-Equivalent groups with Anchor Test (NEAT) design, having both internal and external anchor items (von Davier and von Davier 2004). The NEAT design envisions two populations *P* and *Q*, each represented by samples of examinees that take two different tests. The sample from population *P* takes test *Y*, while the sample from population *Q* takes test *X*. Each of these tests contains a subset of common items *V*. This formulation was appropriate in the present context, where the publisher standardization sample was drawn from *P* and the ECLS-B longitudinal sample was drawn from *Q*. The challenge was to identify items in *V* that act as internal anchor items.[8] In the first experiment reported earlier, using publisher item parameter calibrations, all 66 mental and all 59 motor items were effectively placed in *V*, with no remaining items in *Y* or *X*. The *V* item parameters remained fixed, effectively making these the strongest possible anchor items. This experiment revealed that ECLS-B data were substantially inconsistent with publisher item calibrations.

The second experiment was based on concurrent item calibrations obtained using both ECLS-B and publisher data in a single run. This new design is shown in table 4-9, where all common items were placed in *V* and any remaining items in *Y*.[9] Items from *Y*, *V*, and *X* can be calibrated simultaneously by coding item responses that were not observed and remained missing by design as "not

[8] Internal anchor items are items internal to the test that serve to set the scale metric. External anchor item belong to an external test being used to set the scale metric.

[9] The V item set is a subset of the X item set at this stage, awaiting subsequent analysis, whereupon some of the V items will be transferred to X (with no corresponding publisher item in the Y set). The NEAT design anticipates the second stage of analysis, when some V items will have been transferred to X.

presented.” Several features of this design should be noted. The first feature is that standardization sample observations were calibrated concurrently with ECLS-B observations, yielding a new set of item parameters. These are referred to as ECLS-B item calibrations in order to distinguish them from the original set of publisher item calibrations. The second feature is that BSID-II items not administered in ECLS-B (*Y* in the table) had item parameters that remain fixed so that they effectively acted as external anchor items. These items were positioned across the full range of ability and not just at the extremes of the scale. Parameters for these items remained unchanged during item calibration. The third feature of this design is that parameters for BSF-R *V* items were allowed to float until they found their positions in parameter space relative to the *Y* item parameters that remained fixed.

Table 4-9. First Non-Equivalent groups with Anchor Test (NEAT) design, by item sets: 2001–02 and 2003–04

| Population | | NEAT item sets | | | |
|---|---|---|---|---|---|
| | | *Y* | *V* | *X* | Total |
| Mental | | | | | |
| *P* | Publisher | 112 | 66 | † | 178 |
| *Q* | ECLS-B | † | 66 | 0 | 66 |
| Motor | | | | | |
| *P* | Publisher | 52 | 59 | † | 111 |
| *Q* | ECLS-B | † | 59 | 0 | 59 |

† Not applicable.
NOTE: NEAT = Non-Equivalent groups with Anchor Test design; *Y* = external anchor items with fixed item parameters; *V* = internal conditioned items; *X* = other BSF-R items*;* *P* = publisher standardization dataset*;* *Q* = ECLS-B sample.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04; Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Standardization dataset observations drive the equating design since they were primarily scored on strategically positioned BSID-II items whose parameters remained fixed but also on *V* items whose parameters were allowed to float. Collectively, the standardization dataset observations acted as a set of Bayesian priors on the BSF-R item parameters, coaxing these parameters into positions that were consistent with the fixed set of parameters in *Y.*[10] Admittedly, this was a weak equating design because there were no internal anchor items in the ECLS-B that could act as knots to fix the scale. Instead, calibration relied exclusively on the conditioning provided by publisher standardization dataset observations acting as a stabilizing counterweight. The publisher test effectively played the role of an external anchor test.

This design used common item linkages and Bayesian priors with the full set of BSID-II items to calibrate ECLS-B items in a single run using all 22,391 observations from both the ECLS-B and publisher's standardization datasets. The full set of BSID-II mental items was calibrated using 2,938 publisher standardization dataset observations (13 percent of the total), together with 19,117 (87 percent) observations from the ECLS-B sample. This may overstate the importance of the publisher standardization dataset in one sense and understate it in another. If publisher observations *over the age range covered by ECLS-B sample* were considered, then there were only 1,724 publisher observations of comparable age, which is about 8 percent as large as the total number of ECLS-B observations. On the other hand, BSID-II observations had proportionally more weight in the tails of the ECLS-B ability distributions, where there were relatively few ECLS-B items and where the standardization dataset observations were only needed to help calibrate the BSF-R basal and ceiling item sets. While the numbers of observations involved in calibrating the full set of BSID-II motor items were slightly different, the proportions involved were virtually identical. While publisher observations represented 13 percent of the total combined sample, only 8 percent of those observations were of comparable age.

Full BSID-II item sets with 178 mental items and all 111 motor items were used in the concurrent calibration, including many items that were not present in any of the ECLS-B short forms.[11] There were numerous item linkages relating ECLS-B short forms to the backbone of BSID-II items with

[10] Bayesian priors are probability distributions that are used to condition poorly fitting parameters during estimation. The distributions impose a penalty on improbable parameter values. For the analyses described here, instead of imposing Bayesian priors on individual IRT item parameters, well-conditioned publisher standardization dataset observations were added to the ECLS-B sample during item calibration to accomplish this same purpose. In this role, standardization dataset observations condition the full set of data.

An alternative approach to using observations from the standardization data set would have been to estimate a fully Bayes model with informative prior distributions. When prior distributions are based on publisher item parameter estimates, progressively stronger priors yield parameter estimates that look increasingly more like those obtained with the publisher's dataset. However, the goal of this calibration was to identify a subset of ECLS-B items that were consistent with their publisher counterpart items so that a consistent scale metric could be obtained. This goal was better accomplished with the data augmentation approach described here. Publisher-ECLS-B comparisons subsequently identified a subset of ECLS-B items that were consistent with the corresponding publisher items. These items then became the only direct link between ECLS-B and the publisher in the second stage of this analysis, when item parameters were again estimated, effectively setting the scale.

[11] Conceivably, some of the BSID-II items were too easy for the ECLS-B population and could possibly have been left out of the concurrent calibrations. However, there was no harm done by including these very easy items in the calibrations since their item parameters remained fixed. Items in the extremes of each scale play a much more limited role as external anchor items.

fixed parameters. By using the full set of BSID-II items, it was possible to see how ECLS-B items line up with publisher items across nearly 20 population standard deviations of ability between 1 and 42 months of age. Nor did this exhaust the benefits of this design. It was also possible to separately score the standardization sample observations alternately using either publisher or ECLS-B item calibrations. By scoring the same observations twice with different sets of item parameters, the resulting scale score distributions can be compared, showing the extent to which ECLS-B item parameters replicated the results obtained with the full set of publisher items.

Fit indices for ECLS-B observations scored with the new set of ECLS-B item parameters are reported in table 4-10. All fit indices fell well below the critical value of 1.1. The data fit the IRT model exceptionally well. All indices fell below unity, reflecting the redundancy of information found in assessor-imputed item responses. In general, ECLS-B data were consistent with the item response model. In particular, Outfit indices show that the problem with inconsistent responses far removed from the child's ability level had now been resolved. However, a certain amount of redundant information became apparent throughout, but was most expressive in the motor scale at 9 months. This redundancy was no great cause for concern, because it does not affect scoring, but Infit < 1 implied that IRT standard errors would be somewhat underestimated.

Table 4-10 shows that ECLS-B data fit the response model obtained with concurrent calibration using this first equating design. Sample frequencies reported in table 4-11 confirm this, where virtually all of the observations on both occasions exhibit person fit that was either excellent or acceptable. The improvement in the mental scale on both occasions was most striking. On the motor scale, person fit improved substantially at both 9 months and 2 years. There seems to be little question that the concurrent item calibrations succeeded in producing response models that were consistent with the ECLS-B data.

Table 4-10. Mean fit indices for ECLS-B observations scored after concurrent calibration, by BSF-R scale and round of data collection: 2001–02 and 2003–04

| ECLS-B subsample | Mean squared residual fit index | BSF-R mental item subset | BSF-R motor item subset |
|---|---|---|---|
| 9 months | Information-weighted mean squared residual goodness of fit—Infit | 0.961 | 0.859 |
| | Outlier-sensitive mean squared residual goodness of fit—Outfit | 0.943 | 0.837 |
| 2 years | Information-weighted mean squared residual goodness of fit—Infit | 0.939 | 0.919 |
| | Outlier-sensitive mean squared residual goodness of fit—Outfit | 0.911 | 0.899 |

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Table 4-11. Number and percentage of ECLS-B sample children, by level of fit for the BSF-R scales, after concurrent calibration, by BSF-R scale and round of data collection: 2001–02 and 2003–04

| | | BSF-R mental subset | | BSF-R motor subset | |
|---|---|---|---|---|---|
| Fit | Level of outfit | Number | Percent | Number | Percent |
| 9 months | | | | | |
| Total | 0 ≤ y < ∞ | 10,200 | 100.0 | 10,150 | 100.0 |
| Excellent | 0 ≤ y < 1 | 6,800 | 66.8 | 7,700 | 75.6 |
| Acceptable | 1 ≤ y < 3 | 3,350 | 32.7 | 2,100 | 20.8 |
| Problematic | 3 ≤ y < 5 | 50 | 0.5 | 250 | 2.6 |
| Unacceptable | 5 ≤ y < ∞ | # | # | 100 | 1.0 |
| 2 years | | | | | |
| Total | 0 ≤ y < ∞ | 8,900 | 100.0 | 8,850 | 100.0 |
| Excellent | 0 ≤ y < 1 | 5,950 | 66.9 | 5,800 | 65.8 |
| Acceptable | 1 ≤ y < 3 | 2,950 | 33.0 | 3,000 | 34.2 |
| Problematic | 3 ≤ y < 5 | # | 0.1 | # | 0.1 |
| Unacceptable | 5 ≤ y < ∞ | # | # | # | # |

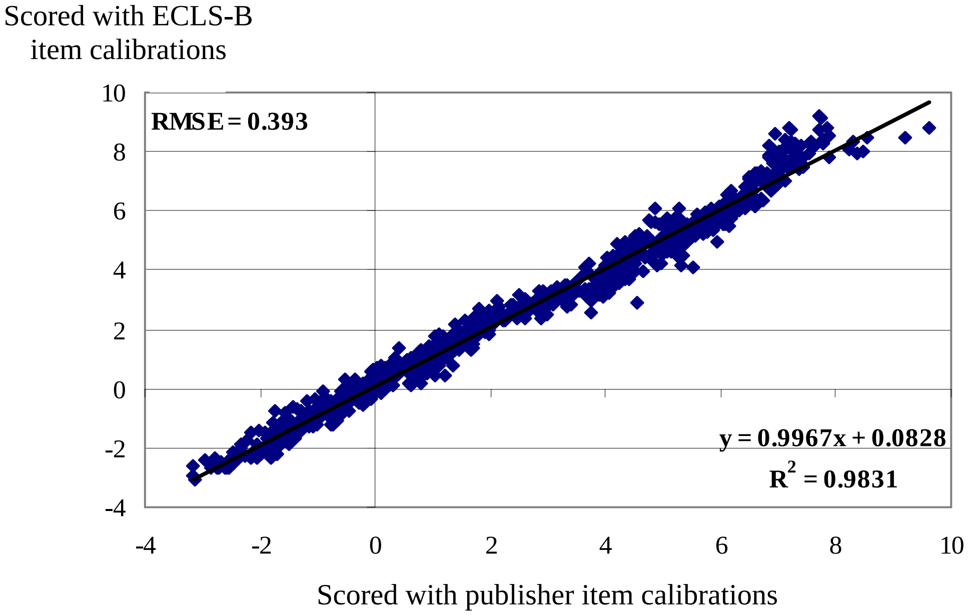
# Rounds to zero.

NOTE: Frequencies may differ slightly from table 4-8 due to weighting and rounding considerations. Outfit = outlier-sensitive mean squared residual goodness of fit. Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

To further examine the quality of item calibrations, it was necessary to consider how the BSF-R scoring performed in relation to the full BSID-II. Figures 4-4 and 4-5 show how the two item sets compared when scoring the same set of 900 standardization sample observations described earlier. The quality of fit was not quite as tight as it was in the previous set of figures, when publisher item parameters were used with both item sets, but it was still respectable. Both $r^2$ coefficients were quite high, although again this was largely a reflection of the enormous range of ability. A better measure of fit was provided by the root mean squared residuals shown at the upper left of the figures, each expressed in population standard deviation units. Generally speaking, the mental scores were accurate to within RMSE = 0.393 of a population standard deviation, while motor scores are accurate to within RMSE = 0.345, when using the full BSID-II as a standard for comparison. In practice, the BSF-R scales were not altogether as precise as publisher item parameters originally had suggested.

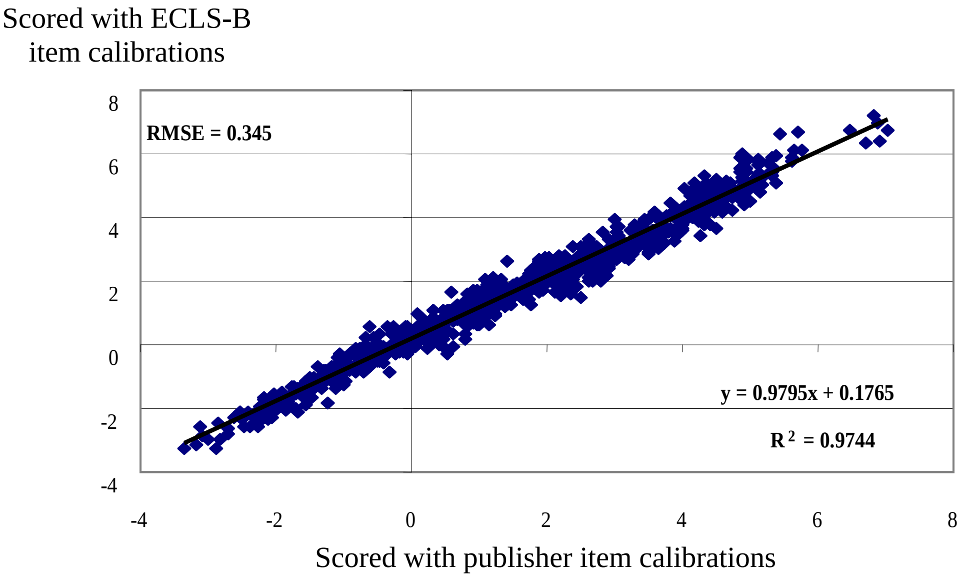
Figure 4-4. Expected a posteriori ability estimates for standardization sample observations scored first with publisher item calibrations (full BSID-II mental items) and then scored with ECLS-B item calibrations (BSF-R) following concurrent item calibration: 2001–02 and 2003–04



NOTE: RMSE = root mean squared error; $R^2$ = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

To the extent that BSF-R item subsets produced results similar to those obtained with the full BSID-II, expected a posteriori (EAP) ability estimates aligned themselves closely with a straight line having an origin of zero and slope of unity.[12] Regression lines in both figures passed close to the scale origin, which coincided with the average ability of 12-month-old infants, and again both slope coefficients were close to unity. Although the BSF-R instruments do not provide anything like τ-equivalent tests, because BSF-R and BSID-II item parameters are often inconsistent, they can still be calibrated on the publisher scale metric.

Figure 4-5. Expected a posteriori ability estimates for standardization sample observations scored first with publisher item calibrations (full BSID-II motor items) and then scored with ECLS-B item calibrations (BSF-R) following concurrent item calibration: 2001–02 and 2003–04



NOTE: RMSE = root mean squared error; $R^2$ = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

## 4.5 BSF-R Equating Design

The new set of ECLS-B item parameters was compared with publisher item parameters so that a set of BSF-R items was identified to serve as internal anchor items. Differential item function (DIF) analysis was used to identify BSF-R items that were inconsistent with publisher item parameters. An item

[12] IRT produces a likelihood function for the response vector at each ability level, $L(X|\theta)$. However, the objective in testing is to obtain an estimate of the probability of an ability given the person's response vector, $P(\theta|X)$. This is known was the expected a posteriori (EAP) probability. Bayes' theorem is used to obtain $P(\theta|X)$, based on the relationship:

$$P(\theta|X) \propto L(X|\theta)\, P(\theta).$$

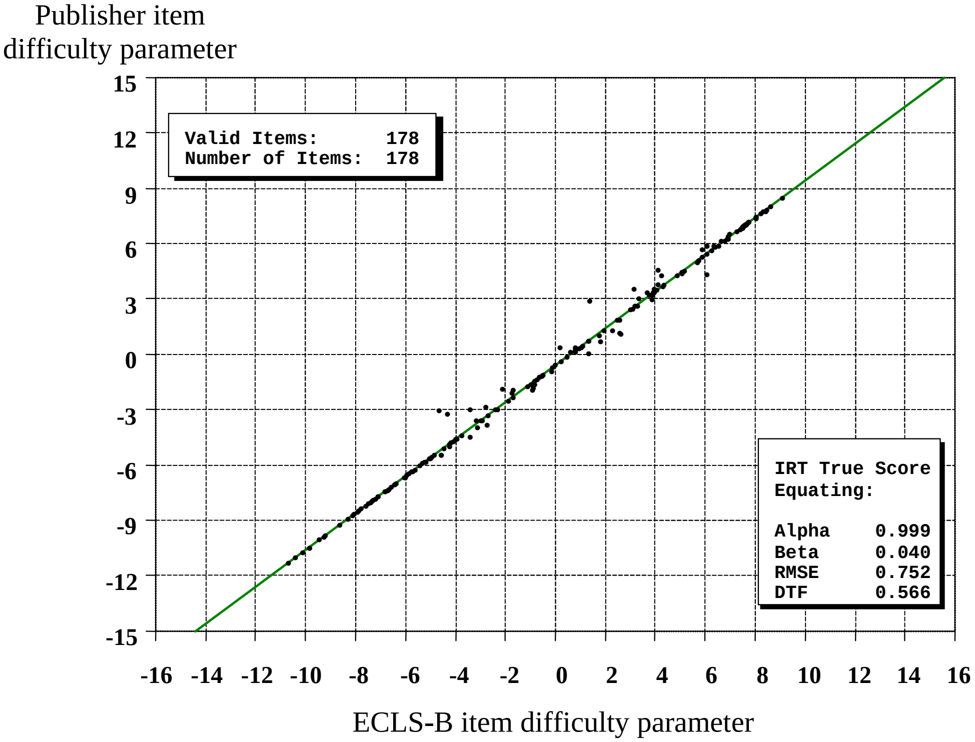
Maximum likelihood is found at the point where this function peaks, also known as the EAP ability estimate. In this sense, the EAP is simply the best available estimate of the person's ability.

has been said to exhibit DIF "if individuals of the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, Swaminathan, and Rogers 1991, p. 110). In the present context, DIF was used in a somewhat different sense to investigate differences between observers and settings rather than differences between population subgroups.

DIF analysis was used to compare BSF-R instruments used by laymen as part of a household survey interview with use of the full BSID-II by trained professionals in a clinical setting. One looks for DIF affecting individuals of the same ability in two populations *P* and *Q*, respectively represented by ECLS-B and the publisher standardization samples. At issue was whether any of the BSF-R items behaved substantially differently in the ECLS-B than they did in BSID-II. The item might still be used in scaling and scoring, but it would play no further role in setting the scale metric. Where this was found to be true, the item was not used to equate the BSF-R with the BSID-II. DIF analysis was used to identify inconsistent items and a subset of highly consistent items that could serve as internal anchor items.

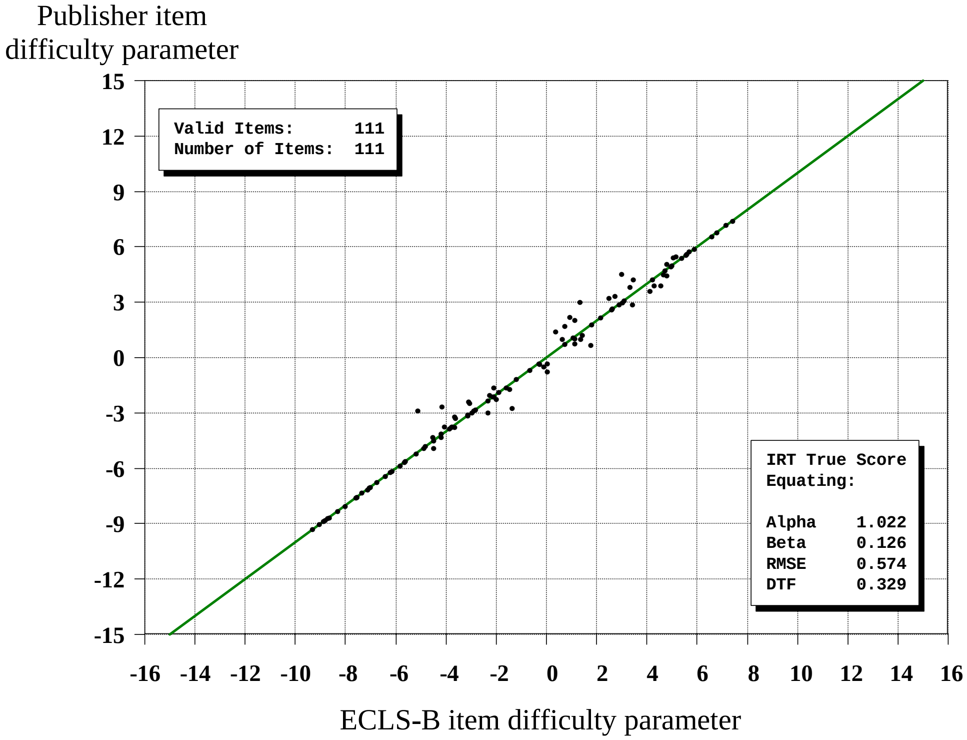
A conceptual grasp of differential item functioning in this context was provided by plotting publisher and ECLS-B item difficulty parameters $b_j$ along perpendicular axes, as shown in figures 4-6 and 4-7. Units of measurement shown in the figures represent population standard deviation units, where publisher observations 12 months of age form the $N(0,1)$ reference population that defines the graph origin and scale. In IRT, item difficulty and person proficiency parameters share a common scale. Bearing in mind this scale metric, there were several instances where the item difficulty parameters of BSF-R items diverge from those of publisher items. This judgment was made empirically to make sure a sufficient number of items would be maintained. These items behaved very differently in the two settings, so much so that they could be considered to be entirely different items lacking any counterpart found among publisher items.

Figure 4-6. ECLS-B mental item difficulty parameters $b_j$ on the *x* axis plotted against the corresponding publisher difficulty parameter on the *y* axis after concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-7. ECLS-B motor item difficulty parameters $b_j$ on the $x$ axis plotted against the corresponding publisher difficulty parameter on the $y$ axis after concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

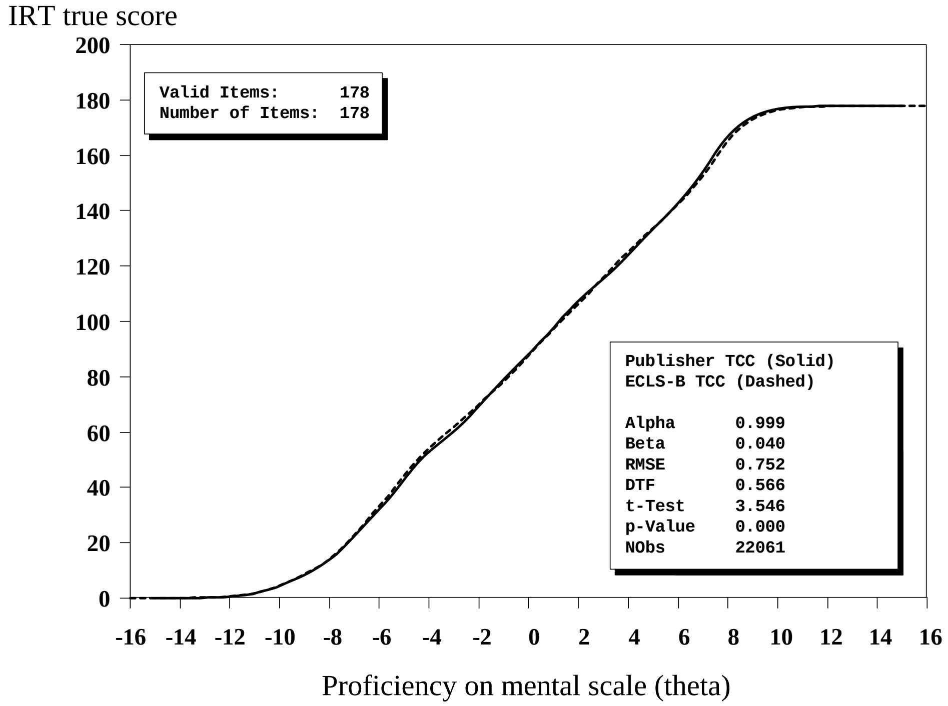
Items showing little or no evidence of DIF have item difficulty parameters $b_j$ located along the diagonal line extending from the lower left to upper right of each figure. Obvious examples include difficulty parameters for *Y* items from the equating design, whose fixed parameters remain the same on both axes. These items are clearly visible lying precisely on the diagonal, especially toward the extremes of the figures, where no BSF-R items are located. At these extremes, BSID-II items appropriate for 1 month of age appear on the diagonal at the lower left of each figure, while BSID-II items appropriate for 42 months appear at the upper right. These item parameter difficulty values appear precisely along the diagonal because item parameters remained unchanged during concurrent calibration. These item parameters convey a clear sense of the central tendency of each scale, which coincides with a 45°-angle line.

Moving closer to the center of each figure, a mixture of BSF-R *V* and BSID-II *Y* items is encountered, where some points lie close to the diagonal while others are farther removed. Here, too, one finds interstitial *Y* items with fixed parameters lying exactly along the diagonal in the midst of other ECLS-B items. Somewhat farther away from the diagonal, one finds item difficulty parameters that are still relatively close to the diagonal. These represent BSF-R items that behaved almost exactly like the corresponding publisher items and were thus worth considering as internal anchor items. Far removed from the diagonal are BSF-R items that are highly inconsistent with the corresponding publisher items. These points represent items that appeared to be much harder or easier in the ECLS-B administration than in the standardization dataset. This inconsistency suggested that these items should play no further role in equating. These items should be considered unique to BSF-R as if they had no counterpart in BSID-II.

The equating constants reported in the box at the lower right of each figure were based on IRT true-score equating (Stocking and Lord 1983). This method used test characteristic curves (TCCs) to align a source test such as BSF-R with a target test such as the publisher. Test equating was accomplished by finding a linear transformation of origin and scale that minimized the weighted area between the two TCCs, as shown in figures 4-8 and 4-9. The equating constants obtained with IRT true-score equating represent the linear transformation that best aligns the two tests. This included a transformation of origin (beta) and of scale (alpha). The figures report values for beta close to zero and values for alpha close to unity. This shows that no true-score equating was required after concurrent calibration.

A measure of differential test functioning (DTF) is the DTF index, where smaller values represent the extent to which the ECLS-B and publisher tests measured the same trait and larger values represent the extent to which the pair of tests fail to align. The DTF index is reported in squared units. The root mean square of this value or RMSE represents the average number-right raw score units separating the two TCCs displayed on the y axis in figures 4-8 and 4-9 (*not* the population standard deviation units displayed on the x axis). The RMSE values in the ECLS-B were both found to be relatively small in relation to the 178 mental and 111 motor items in each respective test. Although the ECLS-B and publisher tests were well aligned over an extensive range of ability, there were many items that were not closely aligned on the two tests. In this case, they should play no further role in *test equating*, although they were retained for scoring.

Figure 4-8. Test characteristic curves (TCCs) for BSF-R and BSID-II mental scales after concurrent calibration: 2001–02 and 2003–04

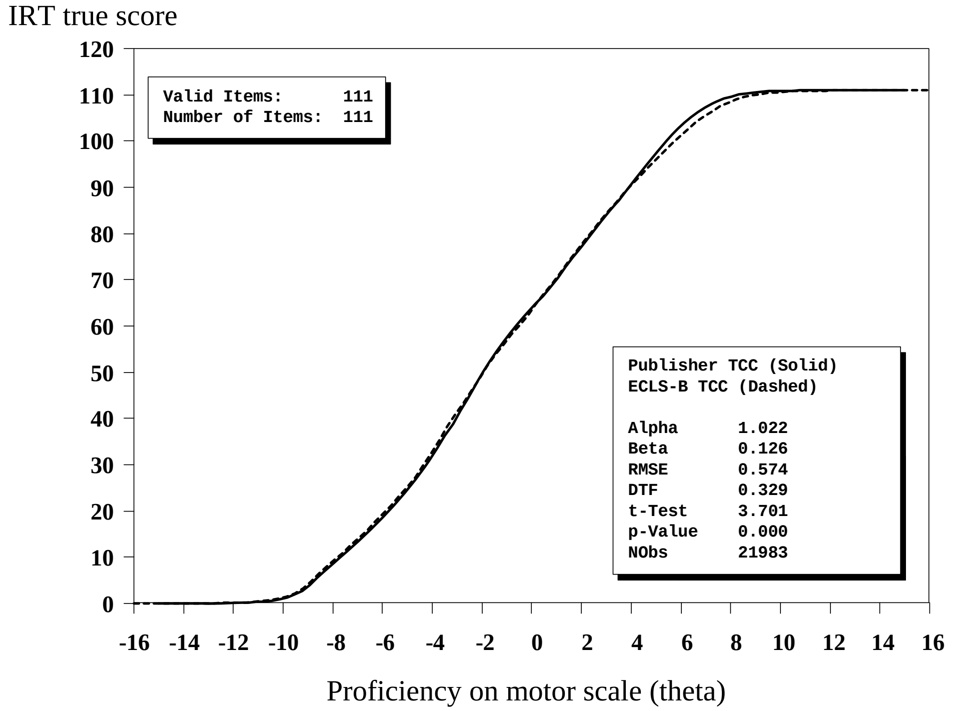


NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Concurrent calibration assures that the two tests are generally well aligned. This alignment is seen both in the diagonal lines shown in figures 4-6 and 4-7, and in the close fit between the two TCCs shown in figures 4-8 and 4-9.[13] In these circumstances, the noncompensatory index (NC-DIF) provides an appropriate measure of individual item DIF (Raju, van der Linden, and Fleer 1995). Parametric IRT models were used to calculate the NC-DIF index. NC-DIF indices represent the weighted mean squared distance between item characteristic curves (ICCs) obtained with separate calibrations. The square-root of the NC-DIF index is thus the weighted average distance separating the two ICCs.

[13] T-test results indicate that there is a significant difference between the TCC based on publisher data and the TCC based on ECLS-B data. However, with large samples such as ECLS-B, relatively small differences will almost always be statistically significant. In this case, the magnitude of the difference is relatively small: 0.752 raw score points on a test that includes 178 items. In practice, children would not be administered all 178 items, but rather only a subset of about 35 items. Out of 35 items, a difference of 0.752 raw score points is 2% to either side of the publisher raw score standard.

Figure 4-9. Test characteristic curves (TCCs) for BSF-R and BSID-II motor scales after concurrent calibration: 2001–02 and 2003–04

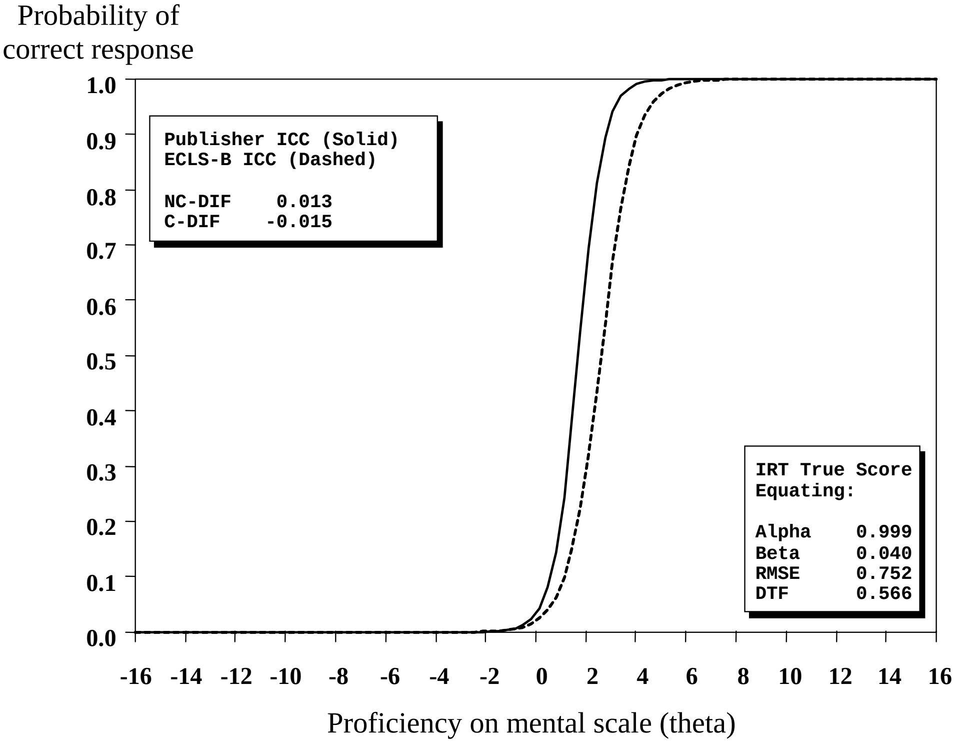


NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

An analysis was conducted to identify BSF-R items exhibiting three different levels of DIF. RMSE < 0.02 were used to identify BSF-R items exhibiting low DIF. These BSF-R items were virtually identical to their counterpart items in BSID-II and were used as internal anchor items with fixed parameters identical to the publisher standard. A mid DIF level with RMSE in the range 0.02 < RMSE < 0.08 was used to identify BSF-R items that played a more limited role in scale equating. These items continued to receive conditioning from publisher standardization dataset observations acting as a stabilizing counterweight. Finally, a residual high DIF level was used for all remaining BSF-R items considered to have no counterpart among BSID-II items. The mean RMSE value in this category was 0.12, implying that ECLS-B and publisher item characteristic curves were fully separated by 12 percentage points.

Figure 4-10 shows an example of an item exhibiting appreciable DIF. Although the two ICCs run broadly parallel to one another, the population-weighted mean vertical distance between them is 0.126 or almost 13 percentage points. Although the ECLS-B item was strongly discriminating, with an item discrimination parameter *a* = 0.856, since it had NC-DIF > 0.08, such items were subsequently disregarded for purposes of scale equating.

Figure 4-10. Item characteristic curves (ICCs) for mental item MEN110 (Names one object) on BSF-R and BSID-II: 2001–02 and 2003–04



NOTE: BSF-R Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; NC-DIF = noncompensatory DIF; C-DIF = compensatory DIF.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

On this basis, ECLS-B items in each of the three DIF levels were identified. Item frequencies for these levels are reported in table 4-12. The table shows that 10 (15 percent) of the mental and 8 (14 percent) of the motor items show virtually no DIF. In all important respects these BSF-R items were identical to their BSID-II counterparts. Another 29 (44 percent) of the mental and 31 (52 percent) of the motor items exhibit tolerable levels of DIF. These items were in the same general vicinity of their BSID-II counterpart items and thus performed as expected. However, 27 (41 percent) mental and 20 (34 percent) motor items performed much differently under household survey conditions than they would have been expected to perform in BSID-II under clinical conditions.

Table 4-12. Frequency count and percentage of BSF-R item parameters (using ELCS-B combined 9-month and 2-year data) for NEAT design items exhibiting low, medium, and high levels of DIF when compared with BSID-II item parameters (using publisher data): 2001–02 and 2003–04

| | | Mental scale | | Motor scale | |
|---|---|---|---|---|---|
| DIF level | NEAT item set | Number | Percent | Number | Percent |
| Total | | 66 | 100.0 | 59 | 100.0 |
| Low | $V_A$ | 10 | 15.2 | 8 | 13.6 |
| Mid | $V_B$ | 29 | 43.9 | 31 | 52.5 |
| High | $X$ | 27 | 40.9 | 20 | 33.9 |

NOTE: NEAT = Non-Equivalent groups with Anchor Test design; $V_A$ = internal anchor items with fixed item parameters; $V_B$ = internal conditioned items; $X$ = other BSF-R items. Detail may not sum to total because of rounding.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Items exhibiting appreciable DIF were examined to see why they may have been sensitive to changes in setting or administrative skill. In a general sense, one could say that many of these items placed excessive demand on interviewer assessors to infer infant intentionality, for example, MEN059—Manipulates Bell, Showing Interest in Detail or Men131—Attends to Story. While it was possible to teach scoring criteria, it was not always possible to teach the interpretation of signs that were sometimes required before a response could be scored. Many ECLS-B interviewers lacked prior experience in child development, and there was no available means to provide them with ready made experience that would enable them to interpret what they observed so as to determine whether a child's response was clearly intentional or not.

The majority of BSF-R items performed in ECLS-B as they would be expected to perform under the best clinical conditions. This too was anticipated since some of the items could be scored objectively, leaving little or no margin for interpretation, for example, MEN089—Puts Six Beads in Box or MEN126—Names Three Objects. The challenge rather was to find a satisfactory procedure for identifying these best performing items.

Results provided by the DIF analysis were used to reformulate the equating design so that BSF-R item calibrations would become more consistent with BSID-II. Those items with appreciable DIF were effectively excluded from equating by reformulating the equating design as shown in table 4-13.

Excluding these items from equating was accomplished within the NEAT framework by transferring BSF-R items with large DIF from item set *V* to *X*. Whereas in the previous concurrent calibration there were no *X* items at all in the design, after this reformulation, there were 27 such items on the mental and another 20 on the motor scale.

Table 4-13. Second Non-Equivalent groups with Anchor Test (NEAT) design: 2001–02 and 2003–04

| Mental | | | | | | |
|---|---|---|---|---|---|---|
| | | NEAT item sets | | | | |
| Population | | *Y* | $V_A$ | $V_B$ | *X* | Total |
| *P* | Publisher | 139 | 10 | 29 | † | 178 |
| *Q* | ECLS-B | † | 10 | 29 | 27 | 66 |

| Motor | | | | | | |
|---|---|---|---|---|---|---|
| | | NEAT item sets | | | | |
| Population | | *Y* | $V_A$ | $V_B$ | *X* | Total |
| *P* | Publisher | 72 | 8 | 31 | † | 111 |
| *Q* | ECLS-B | † | 8 | 31 | 20 | 59 |

† Not applicable.

NOTE: NEAT = Non-Equivalent groups with Anchor Test design; *Y* = external anchor items with fixed item parameters; $V_A$ = internal anchor items with fixed item parameters; $V_B$ = internal conditioned items; *X* = other BSF-R items; *P* = publisher standardization dataset; *Q* = ECLS-B sample.

SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Equating now depended entirely on the remaining 39 (59 percent) mental and 39 (66 percent) motor items exhibiting little or no DIF. A distinction was made between more restricted numbers of virtually identical items that served as internal anchor items $V_A$ and more encompassing numbers of similar but not identical items that continued to receive conditioning in the form of standardization dataset observations acting as the stabilizing counterweight $V_B$. These were the items shown to be most consistent with their respective counterpart publisher items.[14]

Based on this newly reformulated NEAT design, another concurrent calibration was performed, where BSF-R items $V_A$ acted as internal anchors, BSF-R items $V_B$ acted as a stabilizing counterweight and continued to receive conditioning from publisher observations, while the remaining BSF-R items placed in *X* assumed entirely new identities bearing no relation to publisher items. This implied that there were no remaining standardization dataset observations to effectively act as Bayesian priors on the new set of BSF-R *X* items. Standardization dataset responses to these items were removed to *Y*, where all item parameters were based on publisher calibrations and remained fixed during concurrent

[14] By augmenting the data with observations from the standardization dataset during item parameter estimation, the reliability of the Mental test (measured by the ratio of true-score variance to total variance) increased from .84 to .98, a 17 percent improvement. The reliability of the Motor test increased from .96 to .97, a 1 percent improvement.

calibration. Parameters for the new set of BSF-R *X* items were left free to float and find their positions in parameter space based only upon their relationships with other ECLS-B item responses.[15] At the same time, a more highly consistent set of BSF-R $V_B$ items were coaxed into position by standardization dataset observations acting as a set of Bayesian priors. Item parameters were obtained with ECLS-B observations combined with well-conditioned standardization dataset observations. After recalibration, the new set of BSF-R item parameters coincided more closely with ECLS-B item responses and also adhered more closely to the publisher scale metric.

Fit indices for ECLS-B observations scored after DIF analysis in the second concurrent calibration run are reported in table 4-14. These indices remained essentially unchanged from the previous calibration. The reformulated design has resulted in trivial improvements to person fit on the mental scale at both 9 months and 2 years of age. There was essentially no improvement to person fit on the motor scale, where all four indices were only a fraction higher than previously. Fit statistics again show that there was a certain amount of redundant information among BSF-R item responses in the ECLS-B, a tendency that proved to be somewhat more apparent in the motor test at 9 months.

Table 4-14. Mean fit indices for ECLS-B observations scored after DIF analysis and concurrent calibration, by BSF-R scale and round of data collection: 2001–02 and 2003–04

| ECLS-B subsample | Mean squared residual fit index | Mental scale | Motor scale |
|---|---|---|---|
| 9 months | Information-weighted mean squared residual goodness of fit—Infit | 0.954 | 0.862 |
| | Outlier-sensitive mean squared residual goodness of fit—Outfit | 0.932 | 0.829 |
| 2 years | Information-weighted mean squared residual goodness of fit—Infit | 0.940 | 0.937 |
| | Outlier-sensitive mean squared residual goodness of fit—Outfit | 0.909 | 0.920 |

SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Sample frequencies reported in table 4-15 support these same conclusions. Mental observations exhibiting problematic fit improved slightly and declined from 171 observations previously to 136 with the current response model. All remaining improvement in fit on the mental scale was trivial, affecting only a handful of observations. On the motor scale, the number of observations with problematic

[15] BSF-R *X* items showing evidence of DIF remain in the scales for scaling and scoring. Item calibrations reveal that BSF-R items fit ECLS-B data appropriately and thus should be considered as part of each scale. Issues of scale content and construct validity provided additional justification for retaining the items in each scale. When scored, these items increase the precision of ability estimates and, ultimately, enhance scale reliabilities. With maximum likelihood estimation, raw scores play no role in IRT scaling and scoring.

fit actually increased at 9 months and 2 years. Observations exhibiting unacceptable fit increased by 23 at 9 months but declined by 2 at 2 years. The reformulated equating design at best produced trivial improvements in model fit at 9 months of age. However, the scale metric was made more consistent with the second recalibration.

Table 4-15. Number and percentage of ECLS-B sample children, by level of fit for ECLS-B observations scored after differential item function analysis and concurrent calibration, by BSF-R scale and round of data collection: 2001–02 and 2003–04

| | | BSF-R mental scale | | BSF-R motor scale | |
|---|---|---|---|---|---|
| Fit | Level of outfit | Number | Percent | Number | Percent |
| 9 months | | | | | |
| Total | $0 \leq y < \infty$ | 10,200 | 100.0 | 10,150 | 100.0 |
| Excellent | $0 \leq y < 1$ | 6,650 | 65.2 | 7,600 | 74.8 |
| Acceptable | $1 \leq y < 3$ | 3,500 | 34.5 | 2,250 | 22.0 |
| Problematic | $3 \leq y < 5$ | 50 | 0.3 | 250 | 2.4 |
| Unacceptable | $5 \leq y < \infty$ | # | # | 100 | 0.8 |
| 2 years | | | | | |
| Total | $0 \leq y < \infty$ | 8,900 | 100.0 | 8,850 | 100.0 |
| Excellent | $0 \leq y < 1$ | 5,900 | 66.3 | 5,750 | 64.7 |
| Acceptable | $1 \leq y < 3$ | 3,000 | 33.6 | 3,100 | 35.2 |
| Problematic | $3 \leq y < 5$ | # | 0.1 | # | 0.1 |
| Unacceptable | $5 \leq y < \infty$ | # | # | # | # |

# Rounds to zero.

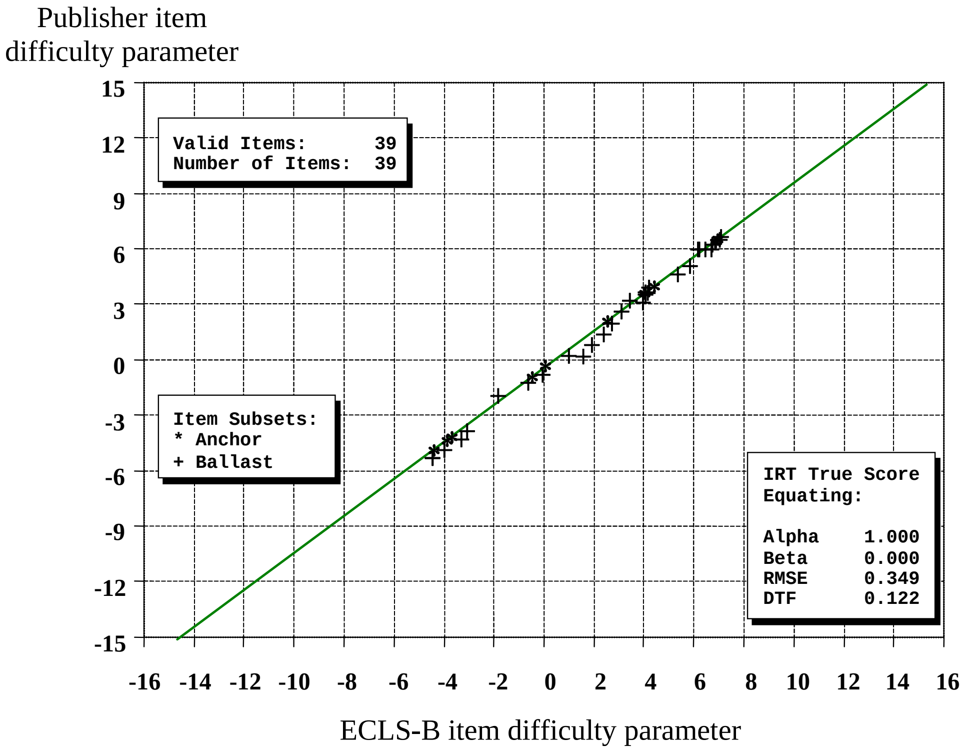
NOTE: Frequencies differ slightly from table 4-8 and table 4-11 due to weighting and rounding considerations. Outfit = outlier-sensitive mean squared residual goodness of fit. Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

As a final step in the item calibrations, publisher items *Y*, which were not used in the ECLS-B, were removed from the BSF-R mental and motor scales. This left only the BSF-R items in each scale, consisting of the $V_A$ and $V_B$ items used in equating and the *X* items that were allowed to float freely. Figures 4-11 and 4-12 show how well the common items *V* performed in equating the BSF-R to the publisher standard. In figures 4-11 and 4-12, ECLS-B item difficulty parameters $b_j$ on the *x* axis are once again plotted against the corresponding publisher parameters on the *y* axis. Publisher items *Y* with fixed parameters, that would otherwise appear on the diagonal, have now been removed from each scale, leaving only the BSF-R items. Notice also that one of the $V_B$ items in the 2-year motor basal item set, MOT062—Walks Alone (Basal), had to be eliminated for the scale due to its dependency with

MOT063—Walks Alone with Good Coordination (Basal), scored not only from a single task administration but also from a single observation, leaving a total of 38 common items in the motor scale.

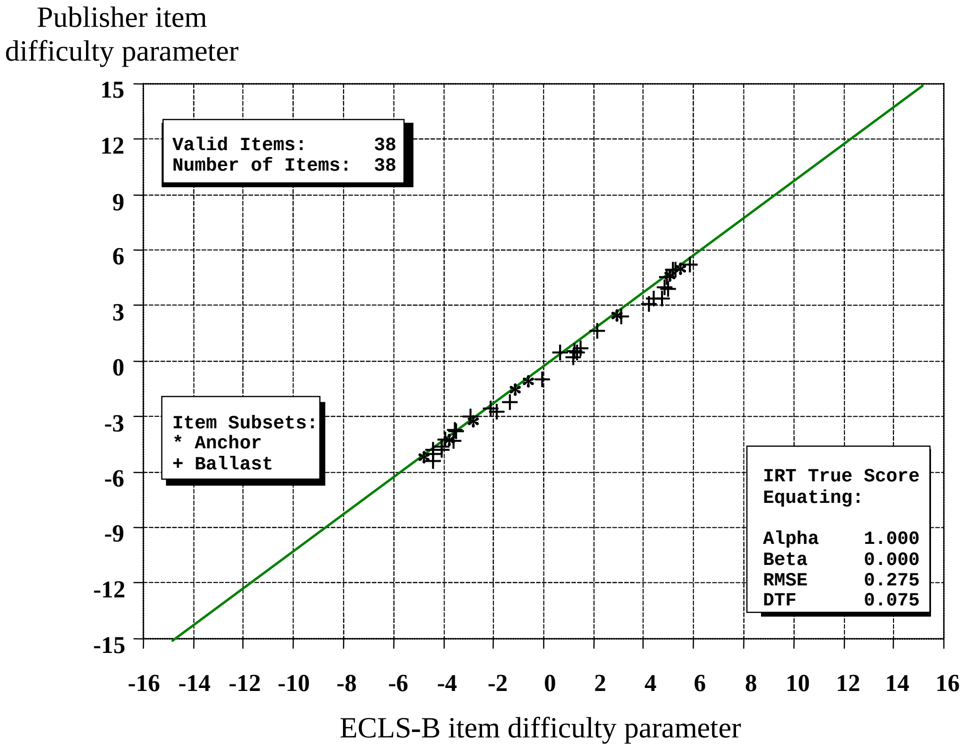
Figure 4-11. ECLS-B mental item difficulty parameters $b_j$ on the x axis plotted against the corresponding publisher difficulty parameter on the y axis after differential item function analysis and concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning. Scatter plot of item difficulties for common items. Anchor items with fixed item parameters lie exactly along the diagonal and stabilizing items conditioned on publisher standardization dataset observations lie near the diagonal.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Item difficulty parameters for BSF-R items used in equating now lie very close to the diagonal in each figure. The two figures show the BSF-R items situated at strategic intervals across a broad range of ability, stretching across approximately 12 population standard deviations on the mental and approximately 11 population standard deviations on the motor scales. These ranges roughly coincide with the 9-month and 2-year latent ability distributions to be shown presently in figures 4-17 and 4-18. These BSF-R items are sufficiently close to the diagonal, sufficient in number, and strategically positioned across a broad range of ability to assure that both the mental and motor scales were effectively calibrated on the publisher scale metric.

Figure 4-12. ECLS-B motor item difficulty parameters $b_j$ on the x axis plotted against the corresponding publisher difficulty parameter on the y axis after differential item function analysis and concurrent calibration: 2001–02 and 2003–04

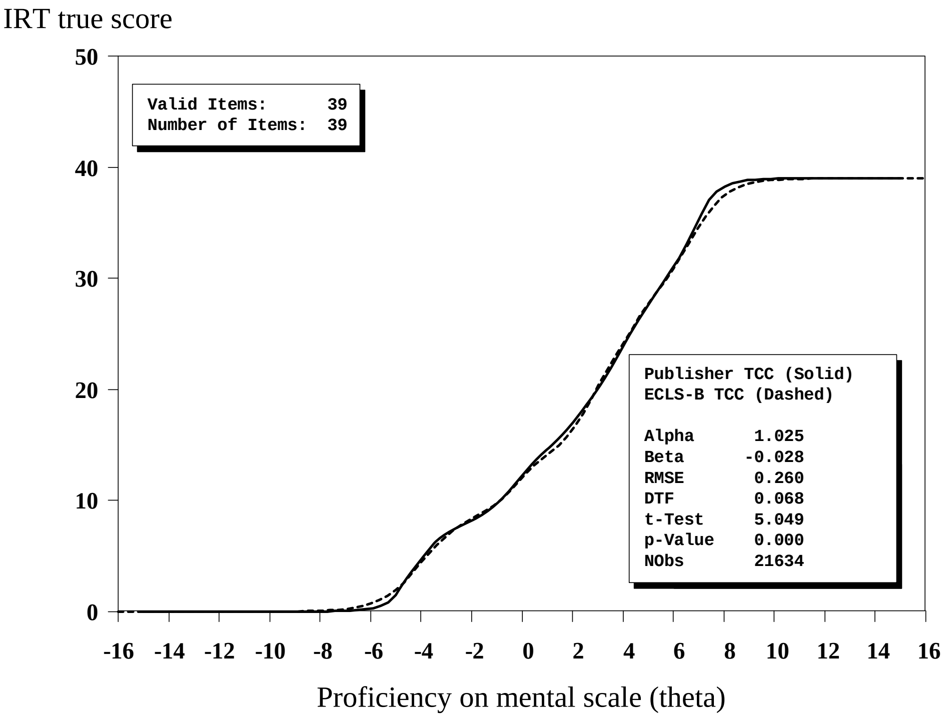


NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning. Scatter plot of item difficulties for common items. Anchor items with fixed item parameters lie exactly along the diagonal and stabilizing items conditioned on publisher standardization dataset observations lie near the diagonal.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-13 shows the TCC alignment for the 39 common items in the mental scale, while figure 4-14 shows the alignment for the 38 common items in the motor scale.[16] The RMSE representing the average distance between the TCCs on the vertical axis for the two tests in both cases is small, although these values need to be considered in relation to the number of items in each test, which has also declined. Although it could be argued that new set of figures represents an improvement in terms of the closer relationship between each pair of TCCs, this ignores the behavior of BSF-R *X* items.

[16] As previously noted, MOT062—Walks Alone (Basal), had to be excluded from the scale due to its dependency with MOT063—Walks Alone with Good Coordination (Basal), leaving a total of 38 common items in the motor scale.

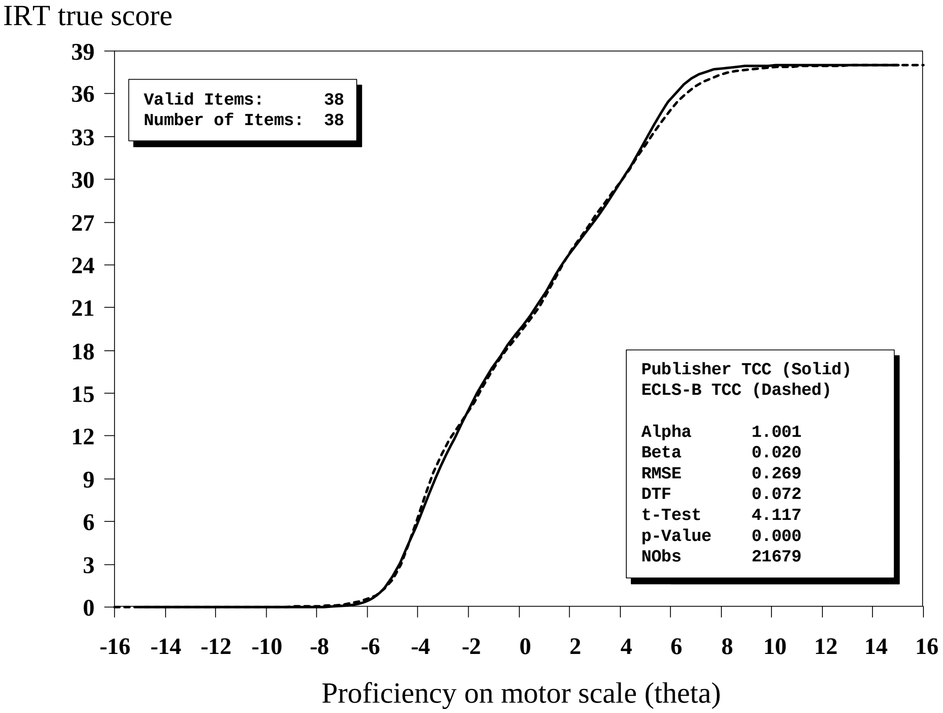
Figure 4-13. Test characteristic curves (TCCs) for BSF-R and BSID-II mental scales after differential item function analysis and concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

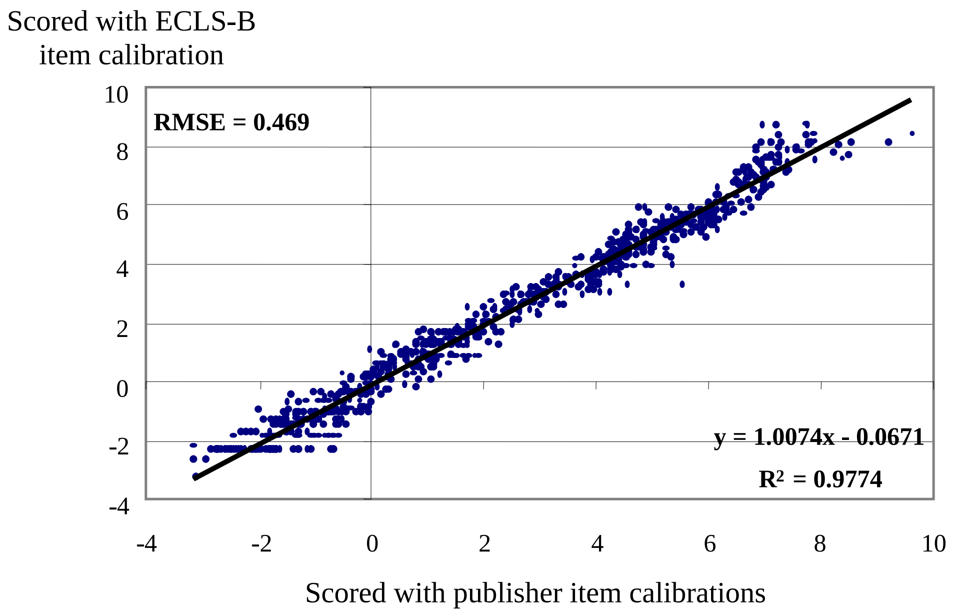
Figure 4-14. Test characteristic curves (TCCs) for BSF-R and BSID-II motor scales after differential item function analysis and concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

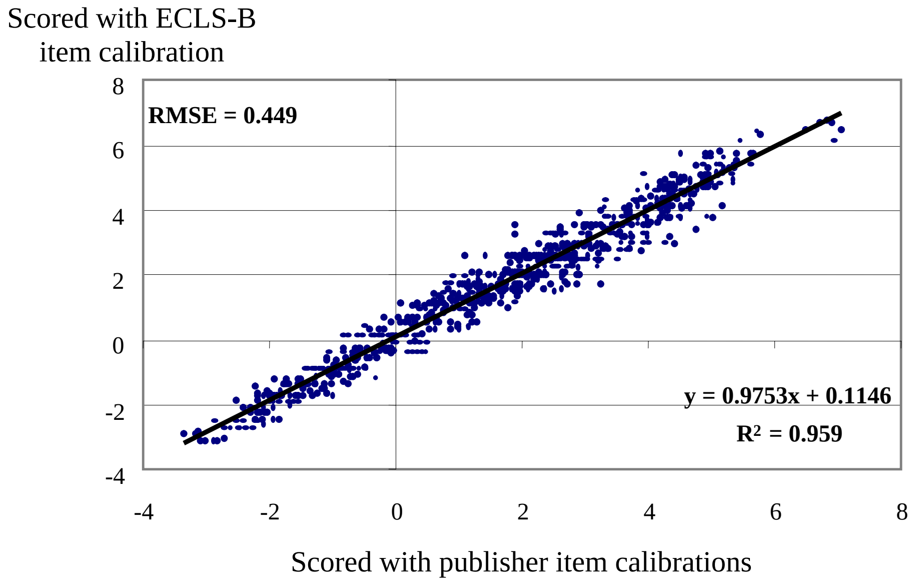
To further examine the quality of item calibrations, consider how BSF-R scoring performed in relation to the full BSID-II on the set of 900 standardization dataset observations. Figures 4-15 and 4-16 show the relationship between the two sets of scores using only the limited number of $V_A$ and $V_B$ BSF-R items when scoring with ECLS-B item parameters. This is because the BSF-R *X* items took on an entirely new identity without parallel in BSID-II. These additional items should have provided some additional precision when scoring ECLS-B observations, but they cannot be used to score publisher observations. The subset of common BSF-R items produced scores that have essentially the same intercept and slope as those produced with publisher item calibrations. Average RMSEs were in the vicinity of 0.45 on both tests. The full set of BSF-R items produced results that were at least as precise and conceivably somewhat more precise.

Figure 4-15. Expected a posteriori ability estimates for standardization sample observations are scored first with the full publisher BSID-II mental items calibrations and then with ECLS-B mental item calibrations (BSF-R) following concurrent item calibration with the second NEAT design: 2001–02 and 2003–04



NOTE: NEAT = Non-Equivalent groups with Anchor Test; RMSE = root mean squared error; $R^2$ = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-16. Expected a posteriori ability estimates for standardization sample observations are scored first with the full publisher BSID-II mental items calibrations and then with ECLS-B motor item calibrations (BSF-R) following concurrent item calibration with the second NEAT design: 2001–02 and 2003–04



NOTE: NEAT = Non-Equivalent groups with Anchor Test; RMSE = root mean squared error; $R^2$ = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

## 4.6 BSF-R Developmental Growth

Infant mental and motor development is so explosive in the early years of life that the range of latent ability found in the ECLS-B spans approximately 12 population standard deviations.[17] Rarely in psychometric research is there an opportunity to capture so much variation in mental and physical status. The challenge for the designers of the BSF-R instruments was to measure each child's mental and motor ability accurately across this broad range of ability, using a reduced item set selected from the BSID-II, while still maintaining comparability with the publisher score metric. After the adjustments discussed above, the design effort met these requirements.

[17] The age-specific latent ability distributions in the publisher standardization dataset have standard deviations that are nearly equal to 1, with small tendency for the variation in mental and physical ability to increase as age approaches 42 months.

Figures 4-17 and 4-18 show kernel density[18] estimations for the ECLS-B mental and motor latent ability distributions for the 9-month and 2-year data collections. Publisher calibrations set the scale metric shown in each of the figures, where standardization data observations in cross-section at 12 months of age have a $N(0, 1)$ distribution, with mean $\mu = 0$ and standard deviation $\sigma = 1$. ECLS-B observations in the 9-month data collection were generally younger than 12 months of age, and thus were represented by negative scale values appearing to the left of each figure. ECLS-B observations in the 2-year data collection were all well above 12 months of age, and were thus represented by positive scale values appearing to the right of each figure.

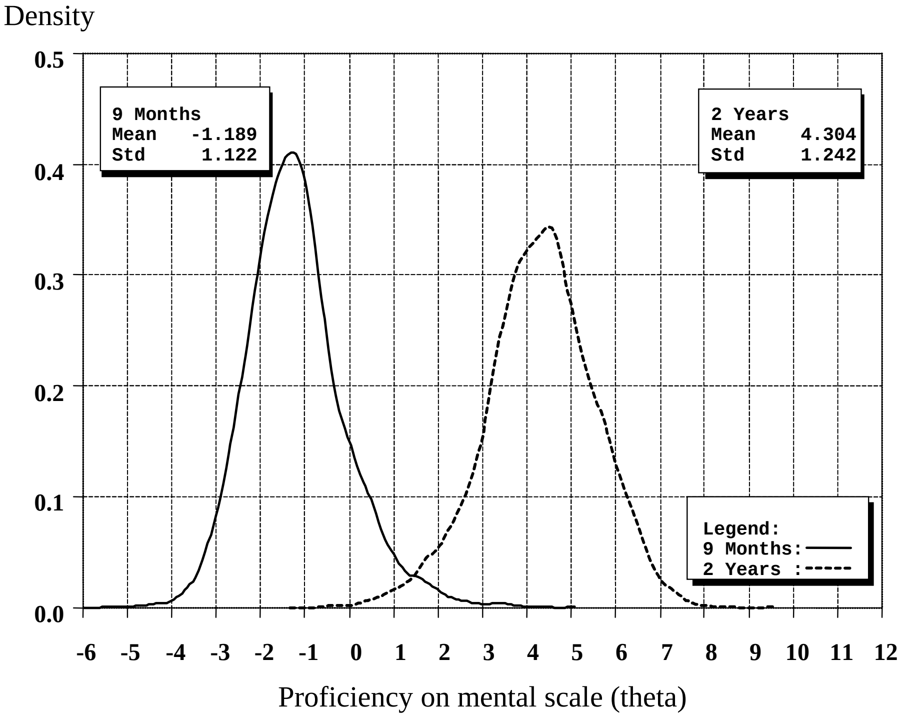
There was considerable variation in mental and motor ability within each cross-section of data, especially for the 9-month data collection. The dispersion of scores is largely a reflection of the distribution of ages in each wave of data. Indeed, the broad range of ages encountered in the ECLS-B implied that mean ability estimates for both the 9-month and 2-year data collections could not reasonably be expected to represent mean ability at exactly 9 or 24 months of age. It was necessary to take this diversity of ages explicitly into account when estimating mean ability at precisely 9 or 24 months.

This could be accomplished by modeling mental and motor scale scores as a function of age at time of assessment. The value of age at precisely 9 months could then be entered into the age-ability equation to obtain a predicted ability score value at precisely 9 months. Indeed, any age value in months could then be entered into the regression equation to predict ability scores across a whole range of ages. In this fashion, a age-ability regression provided a continuous function that could be used to delineate the mean trajectory of ability scores anywhere from say 7 to 28 months, without seriously extrapolating beyond the ages found in the ECLS-B sample, as shown below in figures 4-19 and 4-20.

It should be noted that the ECLS-B is a true longitudinal study, in the sense that the same individuals were assessed at two points in time. Each individual's growth trajectory could thus be summarized in terms of an initial status at exactly 9 months, together with an average monthly growth rate between 9 and 24 months, arriving at a final status at 2 years of age.

[18] A kernel density plot is a non-parametric representation of density that has been smoothed (e.g., by using a Gaussian function).

Figure 4-17. Kernel density estimation for ECLS-B mental latent ability distributions for 9-month and 2-year data collections, in publisher scale metric: 2001–02 and 2003–04

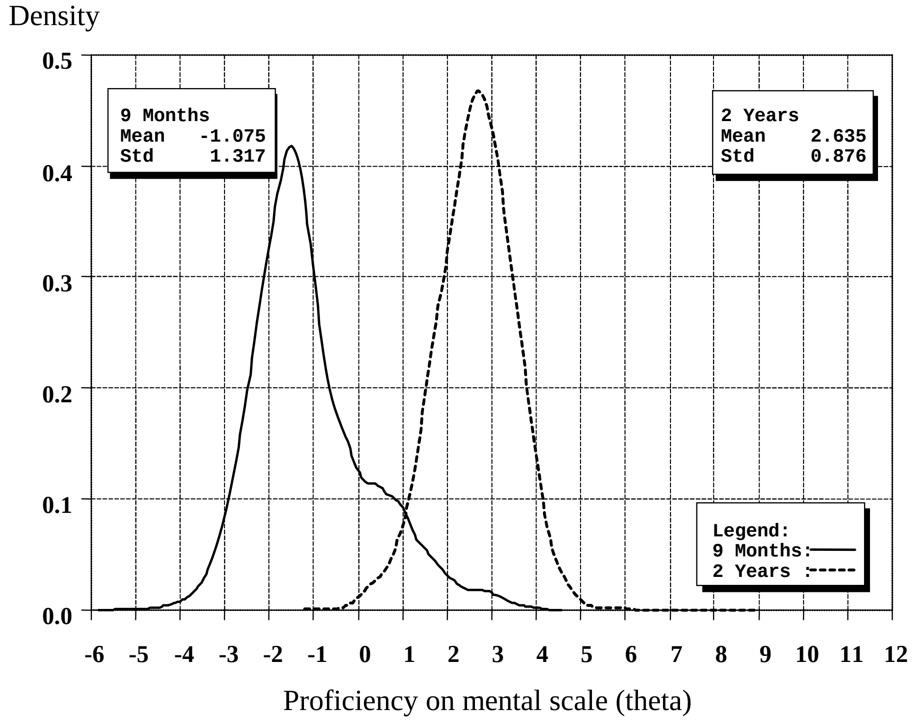


NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

However, as clearly shown in the publisher standardization sample, mental and motor growth is not linear over these ages, and instead decelerates with advancing age. This deceleration introduces a slight curvature in growth trajectories, where growth effectively slows down as age increases. With observations at only two points in time, it was not possible to quantify this degree of curvature for individual growth trajectories in the ECLS-B, however, it was possible to estimate the degree of curvature for the ECLS-B sample as whole.

ECLS-B initial status was not assessed at precisely 9 months of age but over a range of ages extending roughly between 6 and 19 months. Similarly, final status was not assessed at precisely 2 years but over a range of ages extending roughly between 20 and 30 months. Thus, the ECLS-B sample covers a wide age range extending approximately from 9 to 30 months of age. This broad range of ages made it possible to estimate the overall deceleration in mental and motor growth in ECLS-B.

Figure 4-18. Kernel density estimation for ECLS-B motor latent ability distributions for 9-month and 2-year data collections, in publisher scale metric: 2001–02 and 2003–04



NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

The essential comparison of interest was provided by national norms, as represented by the publisher standardization dataset, which was collected in 1991–92. The objective was to determine how the overall mental and motor growth of ECLS-B infants in 2001–03 compared with publisher norms that had been obtained one decade earlier. For this purpose, it should be recalled that the publisher dataset is cross-sectional, in the sense that all infants were assessed at only one point in time. The growth trajectory provided by the publisher standardization dataset is based on a series of successive cross-sections obtained for many ages between 1 and 42 months.

A subset of these observations was selected by the publisher to form a nationally representative sample of infants in order to establish national norms. All of the observations in the standardization dataset were used in regressions modeling mental and motor age-ability relationships reported below, and a dummy variable was included in each regression to identify standardization sample observations. BSID age was used throughout and an age-squared term was used to capture the deceleration in growth.[19]

In the discussion of age-ability regression that follows, decimal BSID months of age rather than chronological age was used throughout. BSID age was obtained from chronological age adjusted for premature birth in ECLS-B. Variables for low birth weight, household socioeconomic status (SES) and race-ethnicity were also included in the ECLS-B regression equations modeling mental and motor status as a function of BSID months of age and a BSID age-squared variable included to capture deceleration in growth in the sample as a whole.

A multilevel analysis was used to model overall mental and motor growth trajectories in relation to age, with time nested within children nested within ECLS-B sample clusters. The multilevel model takes the clustering of the sample design explicitly into account when coefficient standard errors are calculated. The equations for the multilevel, full maximum likelihood model that was estimated are given below.[20] The level-1 model was estimated using 106,450 plausible values, with 5 values per child, usually at two points in time[21]; the level-2 model was estimated using 12,243 children; and the level-3 model was estimated using 159 ECLS-B strata clusters and the 1 publisher group.

Level-1 Model

Y = P0 + P1*(AGE9) + P2*(AGE9SQ) + E

Level-2 Model

P0 = B00 + R0
P1 = B10 + R1
P2 = B20

Level-3 Model

B00 = G000 + G001(PUB) + U00
B10 = G100 + G101(PUB) + U10
B20 = G200 + G201(PUB)

[19] Decimal values for BSID age and the corresponding BSID age-squared variable were centered on zero at 9 months. Publisher sources assure that there were no premature infants in the standardization sample, and standardization sample observations included in the analysis were identified by a dummy variable in the growth analysis regression. The overall intercept coefficient and standardization dummy variable intercept coefficient in the publisher regression were summed together to provide an estimate of the overall mean standardization sample initial status at precisely 9 months of age. The overall slope coefficient and standardization dummy variable slope coefficient in the publisher regression were summed together to provide an estimate the overall mean standardization sample monthly growth rate.

[20] The outcome variable Y is either the Mental or Motor EAP ability estimate. Age9 is months of age centered on exactly 9 months so that the intercept P0 will represent mean ability at precisely 9 months of age. Age9SQ is the squared term for Age9. PUB is the dummy variable used to identify publisher observations belonging to a single group cluster.

[21] Plausible values were used so that the error variances for R0 and R1 could be obtained. Sample weights were divided by 5 to compensate for the number of plausible values per child.

Sample weights, based on sample selection probabilities, were used in the ECLS-B regression. The publisher standardization dataset is self-weighting and represents the national population in 1991–92. The objective here was not to conduct a complete analysis of all these variables but rather to provide an essential summary of the data showing overall mental and motor growth in relation to BSID age. For this purpose, BSID age and the corresponding BSID age-squared variable were centered on zero at 9 months of age.[22]

Figure 4-19 presents the essential summary of the data, where average mental attainment in the ECLS-B sample is compared with average mental ability estimates obtained with the publisher standardization sample. For the ECLS-B sample, the mental age-ability relationship was estimated to be:

$$\theta_{\text{E,Men}} = -1.870 + 0.513x - 0.007x^2,$$

where $\theta_{\text{E,Men}}$ is the mental score obtained with ECLS-B calibrations and $x$ is BSID months of age minus 9 months. Standard errors for each coefficient are small due to the large sample of ECLS-B observations at both points in time. The standard error for ECLS-B mental initial status was 0.028; the standard error for linear growth was 0.013; and the standard error for BSID age-squared was 0.001.

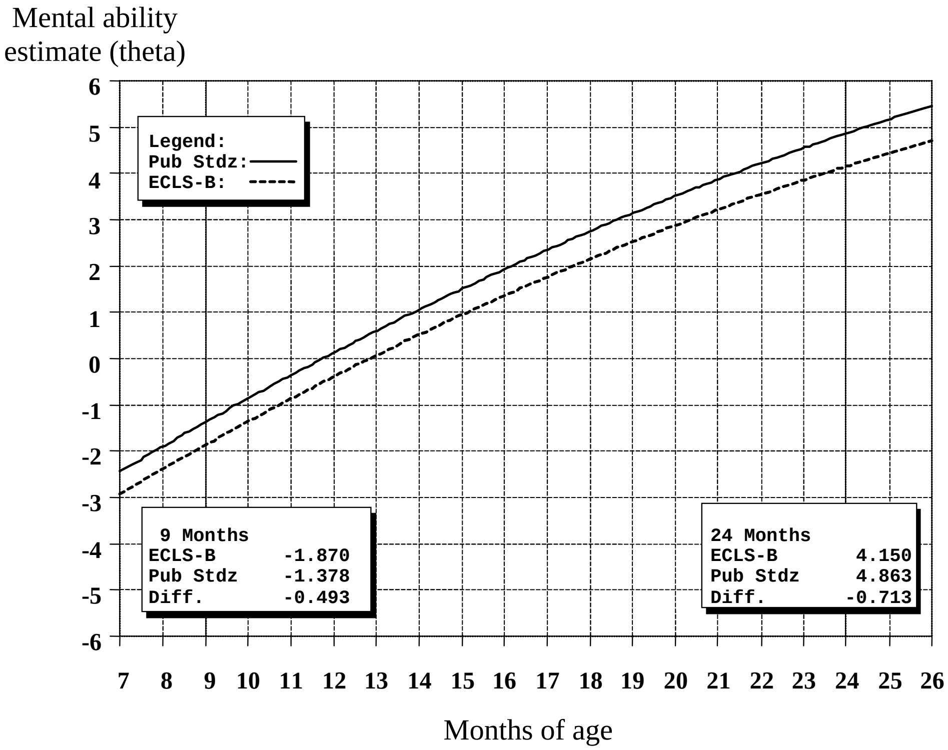
For the publisher standardization sample, the corresponding mental age-ability relationship was estimated to be:

$$\theta_{\text{P,Men}} = -1.378 + 0.521x - 0.007\, x^2,$$

where $\theta_{\text{P,Men}}$ is the mental score obtained with publisher calibrations and $x$ is once again BSID months of age minus 9 months. Standard errors are quite small due to the large number of age cross-sections assessed between 1 and 42 months of age. The standard error for publisher mental initial status was 0.030; the standard error for linear growth was 0.009; and the standard error for BSID age-squared, to three decimal places, was 0.000.

[22] Birth weight was centered at normal weight in the ECLS-B, while household socioeconomic status and each of a series of race-ethnicity dummy variables were centered at their respective ECLS-B sample means. This allowed the ECLS-B regression intercept coefficient to be interpreted as overall mean ECLS-B initial status at precisely 9 months of age among infants who were both carried to term and showed no deficit in birth weight. The ECLS-B regression slope coefficient should be interpreted as the overall mean ECLS-B monthly growth rate for the same population of infants who were both carried to term and showed no deficit in birth weight.

Figure 4-19. Essential summary of the data showing mental growth in relation to age as estimated in publisher standardization sample and ECLS-B sample data: 2001–02 and 2003–04



NOTE: Multilevel regressions obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. “Pub Stdz” refers to the publisher standardization sample; “ECLS-B” refers to the ECLS-B sample.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Conventional tests of statistical significance showed that the difference between the publisher norm and the ECLS-B initial status at precisely 9 months was statistically significant, whereas the difference in monthly growth rate was not statistically significant. To three decimal places, the age-squared coefficients were identical and negative, representing the deceleration in growth over age as measured by the mental scale.[23]

These relationships showed that mental growth was roughly parallel in both the ECLS-B and publisher standardization samples, although the ECLS-B sample initial status revealed an appreciable

[23] These results are reported in the N(0, 1) metric defined at 12 months of age used in IRT scaling and scoring. To translate these results into the *N*(250, 50) metric found in the ECLS-B public-use data files, apply the following formula:

$$\theta_{\text{Men, N(250, 50)}} = [(\theta_{\text{Men, N(0, 1)}} - \overline{X}\ ) / \sigma_x] \times 50 + 250,$$

where $\overline{X}$ = -1.189 represents the ECLS-B mental sample for the 9-month assessment and $\sigma_x$ = 1.124 represents the ECLS-B motor sample standard deviation for the 9-month assessment. A similar transformation would place publisher mental results on the same *N*(250, 50) metric used in the ECLS-B.

deficit in relation to the standardization sample initial status at 9 months. The ECLS-B mean initial status was found at -1.870 and the publisher standardization sample mean initial status at -1.378, where both numbers are expressed in population standard deviations in cross-section at 12 months. The deficit in ECLS-B mean initial status was thus estimated to be -1.870 – (-1.378) = -0.493 or about half a population standard deviation below publisher norms established in 1991–92.

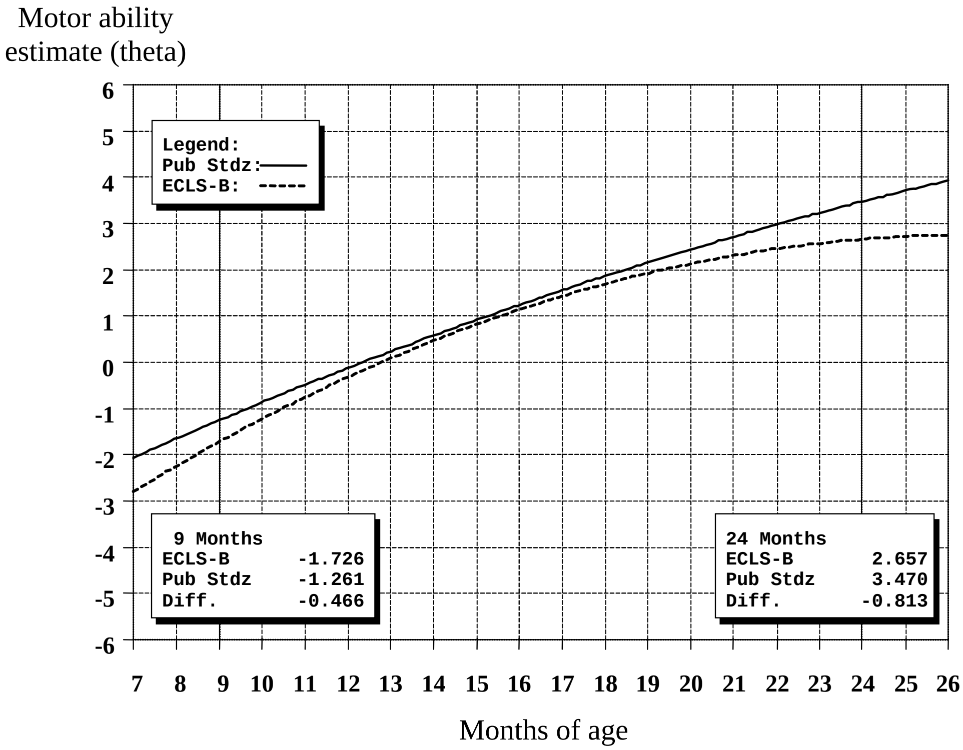
The linear and quadratic components of mental growth were generally similar, and to three decimal places the quadratic terms were identical. However, linear growth in the ECLS-B was found to be somewhat lower than in the publisher standardization sample. Although this growth rate deficit was small, on the order of 0.513 - 0.521 = -0.007 population standard deviations per month, its cumulative effect over 24 - 9 = 15 months would be fairly substantial. In this fashion, the initial ECLS-B deficit of -0.493 at 9 months widened to an estimated deficit of -0.713 population standard deviations at precisely 2 years of age. Not only was the deficit at 2 years statistically significant, it was also relatively large, on the general order of seven-tenths of a population standard deviation. In conducting this analysis, no attempt was made to adjust for demographic differences in the newborn U.S. population. It should be noted that the demographic profile of the infant population changed substantially since 1991–92. During the 1990s, there was substantial immigration resulting in large increases in the Hispanic population. Immigration, combined with high fertility rates among Hispanics, resulted in an increase in the percentage of newborns who were Hispanic. In 1990, 14.5 percent of newborns were Hispanic compared to 23.6 percent in 2001. Not surprisingly, the percentage of newborns with a foreign-born mother also increased from 15.7 percent in 1990 to 22.5 in 2001. Another change was that the percent of births that were preterm rose 12 percent since 1990, from 10.6 percent to 11.9 percent in 2001.

Figure 4-20 presents the essential summary of the data for the motor scale, where the average motor development of the ECLS-B sample is compared with that of the publisher standardization sample. For the ECLS-B sample, the motor age-ability relationship was estimated to be:

$$\theta_{\mathrm{E,Mot}} = -1.726 + 0.511x - 0.015\, x^2,$$

where $\theta_{\mathrm{E,Mot}}$ is the motor score obtained with ECLS-B calibrations and $x$ is BSID months of age minus 9 months. Once again, standard errors are small due to the large size of the ECLS-B sample at both points in time. The standard error for ECLS-B initial status was 0.028; the standard error for linear growth was 0.012; and the standard error for BSID age-squared was 0.001.

Figure 4-20. Essential summary of the data showing motor growth in relation to age as estimated in publisher standardization sample and ECLS-B sample data: 2001–02 and 2003–04



NOTE: Multilevel regressions obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. "Pub Stdz" refers to the publisher standardization sample; "ECLS-B" refers to the ECLS-B sample.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

For the publisher standardization sample, the corresponding motor age-ability relationship was estimated to be:

$$\theta_{\mathrm{P,Mot}} = -1.261 + 0.393x - 0.005\, x^2.$$

Once again, standard errors are small due to the large number of ages assessed in cross-section between 1 and 42 months of age. The standard error for publisher motor initial status was 0.023; the standard error for linear growth was 0.008; and the standard error for BSID age-squared was 0.001.

All three of the differences between the corresponding ECLS-B and publisher coefficients were statistically significant. The deficit in ECLS-B mean initial status relative to the publisher norm at precisely 9 months of age was 11.726 - (-1.261) = -0.466 population standard deviations. ECLS-B linear growth was 0.511 - 0.393 = 0.118 population standard deviations higher than publisher linear growth, but in compensation the ECLS-B coefficient for the BSID age-squared variable was negative and three times as large as the publisher quadratic coefficient. Consequently, both growth trajectories were concave from below, but the ECLS-B curve was more sharply bowed, indicating a higher rate of deceleration in growth. The ECLS-B growth trajectory initially appears to grow more rapidly than the publisher growth trajectory, until the two curves almost intersect in the vicinity of 15 months, after which the ECLS-B curve appears to grow more slowly.[24]

The curvature of the ECLS-B profile appears to be accentuated when this is compared with that of the publisher standardization sample. Despite the broad range of ages found in the ECLS-B sample, this was perhaps insufficient to provide a good estimate of the deceleration in motor growth. However, the ECLS-B sample contained approximately 10,000 observations at 9 months and again at 2 years. For this reason, it is reasonable to assume that initial status at 9 months and the final status at 2 years were accurately estimated. This showed the initial deficit of -0.466 growing to an even larger motor deficit of -0.813 at 2 years or roughly eight-tenths of a population standard deviation. This revealed that the ECLS-B sample of infants in 2001–02 started with an appreciable deficit in initial status at 9 months, growing more slowly on average over the next 15 months, to yield deficit that was almost twice as large at 2 years. The general trend in motor growth resembled that found previously for mental growth,

[24] These results are reported in the N(0, 1) metric defined at 12 months of age used in IRT scaling and scoring. To translate these results into the *N*(250, 50) metric found in the ECLS-B public-use data files, apply the following formula:

$$\theta_{\mathrm{Men,\ N(250,\ 50)}} = [(\theta_{\mathrm{Men,\ N(0,\ 1)}} - \overline{X}\ ) / \sigma_x] \times 50 + 250,$$

where $\overline{X}$ = -1.075 is the ECLS-B motor sample mean for the 9-month assessment and $\sigma_x$ = 1.323 is the ECLS-B motor sample standard deviation for the 9-month assessment. A similar transformation would place publisher motor results on the same *N*(250, 50) metric used in the ECLS-B.

although the curvature of motor growth was found to be much more accentuated. Again, presumably part of the deficit in motor growth should be attributed to demographic changes occurring in the U.S. infant population between 1991–92 and 2001–03.

In a more general sense, this analysis showed that the publisher dataset provided growth coefficient precision comparable to that found in the ECLS-B at a fraction of the sample size by assessing many ages in cross-section. This is clearly an efficient design for establishing national norms. On the other hand, the publisher standardization dataset provided no information about the rate of growth for individuals. Only the ECLS-B sample provided information about the growth rates of individual infants. This is an efficient design for assessing the impact of childrearing and other practices on individual growth rates.

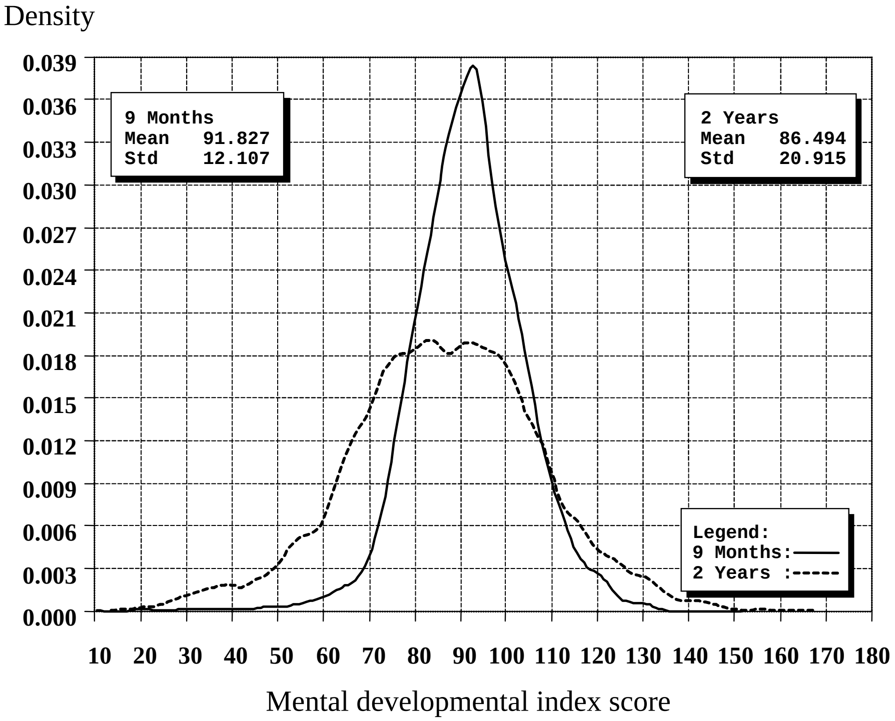
## 4.7 BSF-R Developmental Indices

Fortunately, it is possible to make simple comparisons in cross-section without the need for age-ability regressions. This possibility is provided by publisher developmental index scores. Developmental index scores are age-normed ability estimates. In publisher documentation, developmental index scores are obtained with raw scores and age, adjusted for prematurity, by using a lookup-table. In the ECLS-B, the same scores were obtained by using IRT, BSID age, and regression estimates.

With the benefit of the equating design, EAP ability estimates obtained in the ECLS-B could be reported on the same scale metric used in publisher item calibrations. Thus, EAP ability estimates obtained in the ECLS-B could be used with publisher item calibrations to obtain an IRT true-score, which was a model-based estimate of the publisher's number-right raw score. IRT true-scores, together with BSID age were then applied to a regression equation to produce developmental index scores for individual observations. Figures 4-21 and 4-22 show weighted kernel density estimations for ECLS-B mental and motor developmental index scores.

The developmental index score represents the child's position in an $N(100, 15)$ norm reference distribution with a mean $\mu = 100$ and standard deviation $\sigma = 15$. The two figures show that the developmental status of ECLS-B infants in the weighted sample at the time of the 9-month assessment

Figure 4-21. Weighted kernel density estimations for mental developmental index score distributions in the ECLS-B 9-month and 2-year data collections: 2001–02 and 2003–04

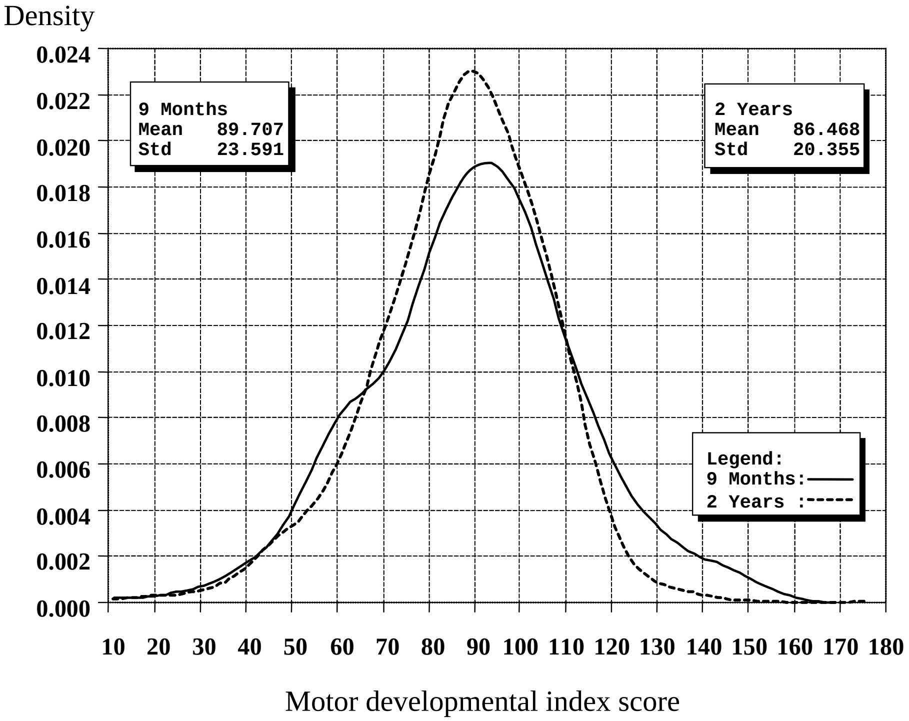


NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

was lower than the national average on both the mental and motor scales. Moreover, by the time of the 2-year assessment, ECLS-B children had fallen farther behind the national standard provided by the publisher on both the mental and motor scales.

The growth deficit on the mental scale was especially dramatic, showing that the central tendency of the ECLS-B sample for the 2-year assessment is nearly a full population standard deviation below norm. The ECLS-B mental distribution was also more heterogeneous—and, therefore, unequal—for the 2-year assessment than it was for the 9-month assessment, with the dispersion in mental developmental index scores almost doubling in size.

Figure 4-22. Weighted kernel density estimations for motor developmental index score distributions in the ECLS-B 9-month and 2-year data collections: 2001–02 and 2003–04



NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

On the motor scale, the ECLS-B sample began with an average deficit disadvantage of two-thirds of a population standard deviation in relation to publisher norms at the time of the 9-month assessment. This relative position declined by the time of the 2-year assessment. By the time of the 2-year assessment, the ECLS-B infant population is almost a full population standard deviation below publisher norms. The variation in motor development decreased very slightly in relation to publisher norms, indicating that the population became relatively more homogenous in terms of psychomotor development.

The central tendencies and the trends in central tendencies over time are similar for both the mental and motor scales. The initial deficits at the time of the 9-month data collection were similar on both the mental and motor scales at approximately two-thirds of a population standard deviation. Both deficits increase to nearly a population standard deviation by the time of the 2-year data collection. By 2 years, ECLS-B found much greater diversity in mental and motor status in 2003–04 compared with the publisher standardization sample obtained in 1991–92. It is possible that the ECLS-B sample as a true probability sample of the U.S. infant population was more inclusive than the publisher norm sample and, therefore, produced lower performance estimates due to broader coverage of the population obtained in the ECLS-B. ECLS-B item responses from 2001–03 consistently reflected performance levels well below those reported by the publisher in 1991–92.

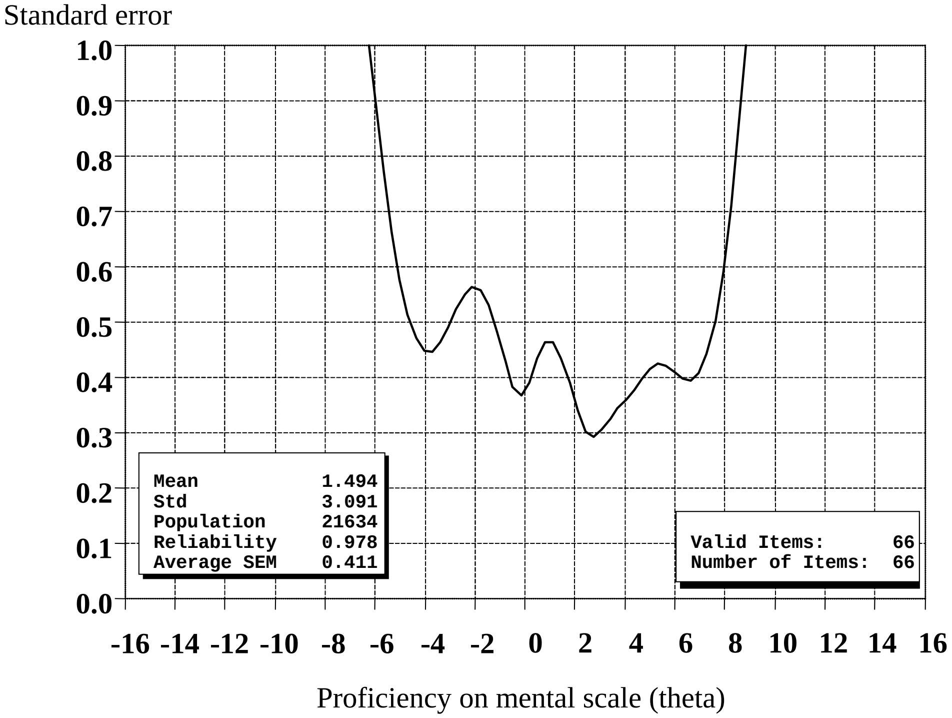
All of these conclusions depend critically on the quality of the ECLS-B equating design. To the extent that scale equating was properly implemented, then these results reflect real differences in the population, and the lower mean developmental scores reflect lower levels of performance across a broad range of developmentally relevant tasks. It is for this reason that the authors of this report have gone to lengths to show that the ECLS-B observations have been scored on the publisher metric. The fact that 60 percent of the mental and 66 percent of the motor items performed the same in the ECLS-B as they do in BSID-II implies that the same scale metric has been maintained throughout. In this case, the mean differences in ability seen among children in the ECLS-B reflect real differences in the infant population, of which the ECLS-B data are representative, rather than artifactual differences that otherwise might be attributable to the BSF-R short form or to fieldwork conditions and procedures used in the ECLS-B.

### 4.8 BSF-R Precision and Reliability

Due to the wide range of age and ability found in the ECLS-B data collections, standard errors of measurement are probably more informative than reliability coefficients as a means for assessing measurement precision. Figures 4-23 and 4-24 show standard errors of measurement at different levels of ability for the BSF-R mental and motor tests used in the ECLS-B. These errors were not adjusted for the redundancy of information observed earlier in BSF-R Infit and Outfit indices because it is not the convention to do so in this type of IRT model. Consequently, standard errors and reliability coefficients reported in the figures and table that follow may appear to be more precise than they actually are. In reviewing these figures, be mindful that the ECLS-B population for the 9-month assessment was centered near $\theta$ = -1.2 on the mental and -1.1 on the motor scale. For the 2-year assessment, the ECLS-B population was centered near $\theta$ = 4.3 on the mental scale and $\theta$ = 2.6 on the motor scale.

The size of standard errors shown in figure 4-23 reveal the limitations in the precision of the BSF-R mental test, where many of the easier items in the 9-month core item set provided little discrimination below $\theta$ = -1.5. Mental basal and ceiling item sets were substantially more informative, providing better precision in the tails of the 9-month distribution. Standard errors over .3 in the core item set indicates a lack of efficiency in basal and ceiling decision rules, implying that some infants failed to receive the required basal or ceiling item sets. The BSF-R mental test generally provided substantially better precision for the 2-year assessment, where precision remained high and standard errors short over most of the latent ability distribution.

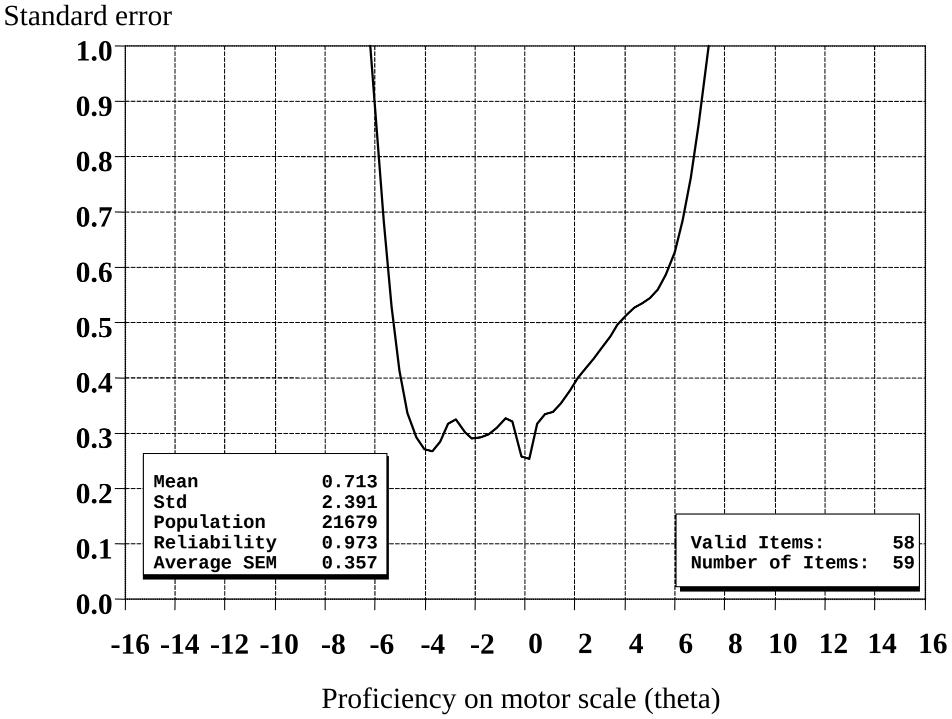
Figure 4-23. Standard errors of measurement for the BSF-R mental test used in the ECLS-B, across all levels of ability: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Std = standard deviation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-24 shows the precision of the BSF motor test across the entire 9-month ability distribution. The situation was different at 2 years, where BSF motor precision was relatively poor at higher levels of ability. Roughly half of the 2-year assessments obtained good to excellent precision before standard errors began to rise substantially at higher levels of ability. The larger standard errors limit reliability in the BSF-R motor test for the 2-year assessment, making this the least reliable of the BSF-R instruments used in the ECLS-B.

Figure 4-24. Standard errors of measurement for the BSF-R motor test used in the ECLS-B, across all levels of ability: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Std = standard deviation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Reliability coefficients reported in the figures were based on normal distributions having the same reported mean and standard deviation. In this sense, these were theoretical reliabilities, based on plausible distributional assumptions rather than directly on ECLS-B observations. From this theoretical perspective, the BSF-R mental was estimated to have an overall reliability of $r_{xx}$ = 0.978 and the motor an overall reliability of $r_{xx}$ = 0.973. The overall IRT reliability coefficient obtained with ECLS-B observations was $r_{xx}$ = 0.975 for the BSF-R mental.[25] The corresponding figure for the BSF-R motor was 0.969. In both cases, the sample-based estimates coincided almost exactly with reliability coefficients calculated based on distribution assumptions.

All of these coefficients were very high because of the broad range of ages considered in the ECLS-B. Standard errors and reliability coefficients by wave of assessment are also reported in table 4-16. These are more realistic reliability coefficients since the differences in age between assessments is no longer a factor. These coefficients show the somewhat lower reliabilities obtained for the mental at the time of the 9-month assessment and the motor at the time of the 2-year assessment. Standard errors continue to be reported in population standard deviation units. The reference population that sets the scale metric is that of the publisher standardization dataset at 12 months of age.

Table 4-16. Standard errors and reliability coefficients for the 9-month and 2-year BSF-R mental and motor-scales: 2001–02 and 2003–04

| Test | Mean standard error[1] | Reliability (internal consistency) |
|---|---|---|
| Mental | | |
| Total | 0.47 | 0.98 |
| 9 months | 0.49 | 0.81 |
| 2 years | 0.44 | 0.88 |
| | | |
| Motor | | |
| Total | 0.38 | 0.97 |
| 9 months | 0.33 | 0.94 |
| 2 years | 0.45 | 0.73 |

[1] Standard errors reported in population standard deviation units, where the 12-month age group is *N*(0, 1).
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

[25] While the information function provides the most comprehensive measure of IRT score reliabilities, it is helpful to provide a single index of test reliability in IRT. For IRT scales, the ratio of the average measurement error variance to total variance can be used for this purpose, after subtracting this value from unity. This yields a measure of true score variance as a proportion of total variance:

$$r_{xx} = \frac{\text{error variance}}{\text{total variance}} \approx 1 - \frac{\sum_{k=1}^{q} \sigma_{e_k}^2 A(X_k)}{\sum_{k=1}^{q} \left(X_k - \overline{X}\right)^2 A(X_k)},$$

where the $A(X_k)$ are normal ordinate weights for points $X_k$ spanning the distribution of ability, with $\sum_{k=1}^{q} A(X_k) = 1$ over $q$ quadrature points.

## 4.9 BSF-R Assessor Effects

This section describes the estimation of assessor effects for repeated measures of mental and motor development used in the ECLS-B.[26] Section 4.6 considered repeated measures of mental and motor status nested within children, nested within sampling clusters. In that section, a multilevel model was used to assess individual growth among children who—for the purpose of that analysis—remained in the same sampling clusters over the course of the investigation. When children cross contextual boundaries during an investigation, the data no longer have such a neat, nested, hierarchical structure. Instead, the analysis involves cross-classifications of children by social settings that change during the course of investigation.

One such migration occurs when considering assessor effects on measures of mental and motor growth used in the ECLS-B, as shown in exhibit 4-1. The data in the table represent only a small selection of ECLS-B assessments for purposes of illustration, broken down by child, data collection, and interviewer. Each row of the table represents a child, whereas each column represents an interviewer. For brevity, only 22 children are listed. The histories of children 050 and 060 illustrate the change in assessor that occurred from the first data collection to the second. These children shared assessor 1020 at the time of the 9-month data collection but were assigned to different assessors at the time of the 2-year data collection, when child 050 was assessed by interviewer 1019 and child 060 was assessed by interviewer 1121. For the 2-year collection, child 060 joined child 150, when both were assessed by interviewer 1121.

Thus, each child assessment could have been conducted by a different interviewer. For a group of children assessed by the same assessor at the time of the 9-month data collection, some of these children might have been assessed by the same interviewer at the time of the 2-year collection, while others were assessed by yet another interviewer. An ECLS-B interviewer might have assessed all of the same children at both points in time but for the 2-year collection could have assessed additional children who were previously assessed by another interviewer. This resulted in a complex data analysis structure, where lower-level units (repeated developmental measures) were cross-classified by two higher-level units (children and interviewer assessors).

Assessor effects on infant growth can be conceived as deflections upward or downward from each child’s individual growth trajectory. In principle, it is possible that measures of mental and motor

[26] A similar estimation of assessor effects in the publisher data could not be made for comparison to the ECLS-B because there is no information about assessors in the standardization dataset.

Exhibit 4-1. Data structure for a selection of ECLS-B assessments, by child, data collection and interviewer assessor: 2001–02 and 2003–04

| | 9-month interviewer assessor | | | | | | | | | | 2-year interviewer assessor | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Child | 1017 | 1019 | 1020 | 1021 | 1023 | 1108 | 1111 | 1166 | 1167 | Child total | 1019 | 1067 | 1108 | 1111 | 1120 | 1121 | Child total | Assessments total |
| Assessor total | 2 | 1 | 3 | 1 | 1 | 8 | 4 | 1 | 1 | 22 | 2 | 3 | 10 | 4 | 1 | 2 | 22 | 44 |
| 050 | | | 1 | | | | | | | 1 | 1 | | | | | | 1 | 2 |
| 060 | | | 1 | | | | | | | 1 | | | | | | 1 | 1 | 2 |
| 070 | | | 1 | | | | | | | 1 | 1 | | | | | | 1 | 2 |
| 150 | | | | | | | | 1 | | 1 | | | | | | 1 | 1 | 2 |
| 180 | | | | | | | | | 1 | 1 | | | | 1 | | | 1 | 2 |
| 190 | | | | | | | 1 | | | 1 | | | | 1 | | | 1 | 2 |
| 260 | | | | | | | 1 | | | 1 | | | | 1 | | | 1 | 2 |
| 310 | | | | | 1 | | | | | 1 | | | | | 1 | | 1 | 2 |
| 380 | 1 | | | | | | | | | 1 | | 1 | | | | | 1 | 2 |
| 390 | 1 | | | | | | | | | 1 | | 1 | | | | | 1 | 2 |
| 400 | | | | 1 | | | | | | 1 | | 1 | | | | | 1 | 2 |
| 440 | | | | | | 1 | | | | 1 | | | 1 | | | | 1 | 2 |
| 450 | | 1 | | | | | | | | 1 | | | 1 | | | | 1 | 2 |
| 460 | | | | | | 1 | | | | 1 | | | 1 | | | | 1 | 2 |
| 470 | | | | | | 1 | | | | 1 | | | 1 | | | | 1 | 2 |
| 550 | | | | | | | 1 | | | 1 | | | 1 | | | | 1 | 2 |
| 601 | | | | | | 1 | | | | 1 | | | 1 | | | | 1 | 2 |
| 602 | | | | | | 1 | | | | 1 | | | 1 | | | | 1 | 2 |
| 610 | | | | | | 1 | | | | 1 | | | 1 | | | | 1 | 2 |
| 640 | | | | | | 1 | | | | 1 | | | 1 | | | | 1 | 2 |
| 700 | | | | | | 1 | | | | 1 | | | 1 | | | | 1 | 2 |
| 730 | | | | | | | 1 | | | 1 | | | | 1 | | | 1 | 2 |

NOTE: Selection of ECLS-B sample observations for an unweighted multilevel regression of age-ability relationships cross-classified by child and interviewer assessor. Each 1 represents a single assessment.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

growth in the ECLS-B were deflected upward or downward by exposure to different assessors. Assessor effects would be represented by the variance of this deflection. Part of the variation in growth curves that otherwise might have been attributed to individual growth differences among infants would instead be attributed to assessor effects. Hence, it was desirable to estimate the magnitude of assessor effects. Taking assessor effects explicitly into account was expected to reduce the temporal instability in infant outcomes.

A cross-classified random effects model, estimated using full maximum likelihood estimation, was used to examine the magnitude of assessor effects on measures of child developmental status and growth used in the ECLS-B. Unweighted ECLS-B data, rather than weighted data, were used because the purpose of the analyses was to evaluate the quality of the ECLS-B assessment data, rather than to generalize the findings to the population of children. HCM2, part of the HLM 6.0 software package, was used for this purpose. ECLS-B age-ability relationships were modeled in much the same way as described previously in section 4.6 of this report. This new model again included BSID age centered at precisely 9 months of age and a corresponding BSID age-squared variable. Continuous variables representing premature birth and household SES were added to the model, as were dummy variables indicating whether a child was Black or Hispanic. The equations for the 2-level model used for the mental scale are specified below.[27] The level-1 model was estimated using 91,430 plausible values, with 5 values per child, usually at two points in time; the level-2 model was estimated using 9,412 children.

Level-1 Model

```
        Y = P0 + P1*(AGE9) + P2*(AGE9SQ) + e
```

Level-2 Model

```
        P0 = theta(0) + b00 + c00
                 + (G01)*PREMONTH
                 + (G02)*SESML
                 + (G03)*HISP

        P1 = theta(1) + b10 + c10
                 + (G11)*PREMONTH
                 + (G12)*SESML
                 + (G13)*AFAM
                 + (G14)*HISP

        P2 = theta(2)
```

[27] The outcome variable Y is the Mental EAP ability estimate. Age9 is months of age centered on exactly 9 months so that the intercept P0 will represent mean ability at precisely 9 months of age. Age9SQ is the squared term for Age9. PREMONTH is months premature at birth. SESML is a maximum likelihood scale score representing socio-economic status. HISP is a dummy variable representing Hispanic and AFAM is a dummy variable representing African-American children.

The independent variables, which are all characteristics of children, used in this analysis were chosen because they are consistently related to status or growth, or both, in ECLS-B. They were not included in order to assess the impact of these variables, because the objective was to examine the extent to which assessor effects deflect measures of mental and motor growth between 9 months and 2 years of age. Premature birth, SES, and ethnicity were accounted for in the model so that none of these factors would be confounded with assessor effects.

In this analysis, the essential summary of mental growth in relation to age was given by the regression equation:

$$\theta_{\text{E, Men}} = -1.872 + 0.459x - 0.004x^2,$$

where $\theta_{\text{E, Men}}$ is child mental scale score and $x$ is BSID age in ECLS-B. The interested reader may want to compare this growth curve with the formula shown earlier: $\theta_{\text{E,Men}} = -1.870 + 0.513x - 0.007x^2$. Standard errors for each of these coefficients were once again small due to the large sample of ECLS-B observations at both points in time. The standard error for ECLS-B mental initial status at precisely 9 months of age was 0.018; the standard error for the linear monthly growth rate was 0.005 and the standard error for BSID age-squared to three decimal places was 0.000. The reader may want to compare these values with those reported earlier in section 4.6. Standard errors for the cross-classified mental regression coefficients were in every case smaller than those found in the original age-ability equation reported in section 4.6 of this report.

Part of the variability that had once been attributed to individual child differences could now be attributed to assessor effects. Additionally, some of the within child variation in growth could also now be attributed to assessor effects. The intra-class correlation for mental initial status at precisely 9 months of age, conditional upon growth and the additional child control variables (premature birth, SES and ethnicity), was given by the ratio of the variance for child initial status in relation to total variance:

$$\frac{\tau_{b_{00}}}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.386}{0.386 + 0.040 + 0.235} = 0.584,$$

or about 58 percent of total variance, where $\tau_{b_{00}}$ is between children true score variance, $\tau_{c_{00}}$ is between assessor variance, and $\sigma^2$ is random error variance. This value can be directly interpreted as a reliability coefficient, defined in classical test theory as the ratio of true score variance to total variance. From this perspective, the value of 0.584 represents the reliability of the BSF-R measure of mental status in cross-section at precisely 9 months of age. The complement to this value is 1 - 0.584 = 0.416 or about 42 percent, reflecting relatively high levels of measurement error of one form or another on the BSF-R mental test at 9 months (random error plus assessor effects).[28]

The proportion of random error in relation to total variance on the mental test was:

$$\frac{\sigma^2}{\tau_{b_{00}}+\tau_{c_{00}}+\sigma^2}=\frac{0.235}{0.386+0.040+0.235}=0.355,$$

or roughly 36 percent of the total variance in initial status at precisely 9 months of age. Assessor effects on initial status thus appeared to be relatively small partly because random error was so large. On the BSF-R mental test, random error accounts for more than a third of the total variance in initial status at 9 months.

The intra-class correlation for assessor effects on mental initial status is given by the ratio of assessor variance to total variance:

$$\frac{\tau_{c_{00}}}{\tau_{b_{00}}+\tau_{c_{00}}+\sigma^2}=\frac{0.040}{0.386+0.040+0.235}=0.061,$$

which represents about 6 percent of total variance in cross-section at precisely 9 months of age. Thus, the impact of assessor effects on the measurement of mental initial status was small, representing only 6 percent of total variance. As a proportion, assessor effects on mental measures represented 0.040 / 0.275 = 0.146 or about 15 percent of total measurement error (assessor effects plus random error). The complement to this figure was 1 - 0.146 = 0.854 or about 85 percent, which represented random error, by far the largest component of total measurement error found in the BSF-R mental. From this perspective, assessor effects on mental initial status appeared to be relatively small if only because random error was so large.

Variance components as usual were reported in squared units of measurement. The corresponding standard deviations for these values were $\sigma_{b_{00}}=\tau_{b_{00}}{}^{1/2}=\sqrt{0.386}$ = 0.621 for mental true score initial status; $\sigma_{c_{00}}=\tau_{c_{00}}{}^{1/2}=\sqrt{0.040}$ = 0.200 for assessor effects on mental initial status; and $\sigma=\sqrt{0.235}$ = 0.484 for the random error in initial status. These units were the same as those used in scaling and scoring the BSF-R mental test, expressed in standard deviation units of the 12-month-old cohort found in the publisher standardization dataset.

[28] Five plausible values were used to represent internal inconsistency measurement error at each point in time for each observation.

Turning to consider the impact of assessor effects on mental monthly growth rate, the variance attributed to assessor effects was $\tau_{c_{10}}$ = 0.001, while true score growth rate variance among children was $\tau_{b_{10}}$ = 0.006. With data at only two points in time, the random error component for growth rates could not be estimated. Instead, the importance of assessor effects on growth rates was obtained by considering the variance of assessor effects in relation to the true-score variance of growth rates, which was estimated to be $\tau_{c_{10}} / \tau_{b_{10}}$ = 0.001 / 0.006 = 0.105 or about 11 percent as large as the true-score variance in mental growth rates.

In standard deviation units, this was $\sigma_{c_{10}} = \tau_{c_{10}}{}^{1/2} = \sqrt{0.001}$ = 0.033 or assessor effects on mental growth rates and $\sigma_{b_{10}} = \tau_{b_{10}}{}^{1/2} = \sqrt{0.006}$ = 0.078 for true score variation in growth between children. How should one interpret the size of assessor effects in relation to growth? The average growth rate was estimated to be 0.459 population standard deviations per month, so the expected impact of an assessor effect one standard deviation above average mental growth would be 0.459 + 0.033 = 0.492 units per month, whereas the impact of an assessor effect one standard deviation below average growth would be 0.459 - 0.033 = 0.427 population standard deviations per month. The difference in mental growth rates over these two extremes would be 0.492 - 0.427 = 0.065 of a population standard deviation per month.

A parallel HCM2 analysis was used to obtain an essential summary of motor growth in relation to age in ECLS-B, as represented by the regression equation:

$$\theta_{\text{E, Mot}} = -1.742 + 0.516x - 0.015x^2,$$

where $\theta_{\text{E, Mot}}$ is the child motor scale score and $x$ is once again BSID age. The interested reader may want to compare this growth curve with the equation shown earlier: $\theta_{\text{E,Mot}} = -1.726 + 0.511x - 0.015\ x^2$. Standard errors were again small due to the large size of the ECLS-B sample. The standard error for ECLS-B motor initial status at precisely 9 months of age was 0.022; the standard error for the linear monthly growth rate was 0.005 and the standard error for BSID age-squared to three decimal places was 0.000. Standard errors for the cross-classified random effects model of assessor effects were again smaller in every case than those reported for the original motor regression equation in section 4.6 of this report.

The intra-class correlation for motor initial status at precisely 9 months, conditional upon growth and the additional child control variables (premature birth, SES, and ethnicity), was given by the ratio of variance of child initial status to total variance:

$$\frac{\tau_{b_{00}}}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.918}{0.918 + 0.053 + 0.155} = 0.815,$$

or about 82 percent of total variance, due to the higher reliability of the BSF-R measure of motor status at precisely 9 months of age. The complement to this value is $1 - 0.815 =$ 0.185 or about 19 percent, reflecting relatively low levels of measurement error of one form or another on the BSF-R motor test at 9 months (random error plus assessor effects).

The intra-class correlation for assessor effects on the BSF-R motor is given by the ratio of assessor variance to total variance:

$$\frac{\tau_{c_{00}}}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.053}{0.918 + 0.053 + 0.155} = 0.047,$$

which represents about 5 percent of total variance in cross-section at precisely 9 months of age. Thus, the impact of assessor effects on the measurement of motor initial status was again small, representing only 5 percent of total variance.

As a part of total measurement error (assessor effects plus random error), assessor effects on the motor represented 0.053 / 0.208 = 0.255 or about 26 percent of total measurement error. The complement to this was 1 - 0.255 = 0.745 or about 75 percent, which represented random error found in the BSF-R motor test, again by far the largest component of total measurement error.

The proportion of random error in relation to total variance was:

$$\frac{\sigma^2}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.155}{0.918 + 0.053 + 0.155} = 0.138,$$

or roughly only 14 percent of total variance. Assessor effects on motor initial status at 9 months appeared to be relatively more substantial if only because random error was so much smaller on the BSF-R motor. On the BSF-R mental, random error represented fully 36 percent of the total variance in initial status whereas on the motor random error represented only 14 percent of the total variance in initial status.

Variance components as usual were reported in squared units of measurement. The corresponding standard deviations for these values are $\sigma_{b_{00}} = \tau_{b_{00}}{}^{1/2} = \sqrt{0.918}$ = 0.958 for motor true score initial status; $\sigma_{c_{00}} = \tau_{c_{00}}{}^{1/2} = \sqrt{0.053}$ = 0.230 for assessor effects on initial status; and $\sigma = \sqrt{0.155}$

= 0.394 for the random error in initial status. These units were the same as those used in scaling and scoring the BSF-R motor test, which again were based on the motor standard deviation for the 12-month-old cohort found in the publisher standardization dataset.

Turning to consider the impact of assessor effects on motor monthly growth rate, the variance attributed to assessor effects was $\tau_{c_{10}}$ = 0.001, while true score growth rate variance among children was $\tau_{b_{10}}$ = 0.006. Although, to three decimal places, these values appeared to be identical to corresponding values reported previously for mental growth, in fact, these values were smaller by just a tiny fraction. The importance of assessor effects on growth rates was obtained by considering the variance of assessor effects in relation to the true-score variance of growth rates, which was estimated to be $\tau_{c_{10}} / \tau_{b_{10}}$ = 0.001 / 0.006 = 0.114 or again about 11 percent as large as the true-score variance in motor growth rates.

In standard deviation units, this variation was $\sigma_{c_{10}} = \tau_{c_{10}}{}^{1/2} = \sqrt{0.001}$ = 0.025 for assessor effects on motor growth and $\sigma_{b_{10}} = \tau_{b_{10}}{}^{1/2} = \sqrt{0.006}$ = 0.074 for true score variation in motor growth between children. The importance of assessor effects can again be assessed in relation to average growth. The average motor growth rate was estimated to be 0.516 population standard deviations per month in cross-section, so the expected impact of an assessor effect one standard deviation above average motor growth would be 0.516 + 0.025 = 0.541 per month, while the impact of an assessor effect one standard deviation below average growth would be 0.516 - 0.025 = 0.491 population standard deviations per month. The difference in motor growth rates over these two extremes would be 0.541 –0.491 = 0.050 of a population standard deviation per month. The value of this difference was somewhat smaller for motor than it was for mental growth rates.

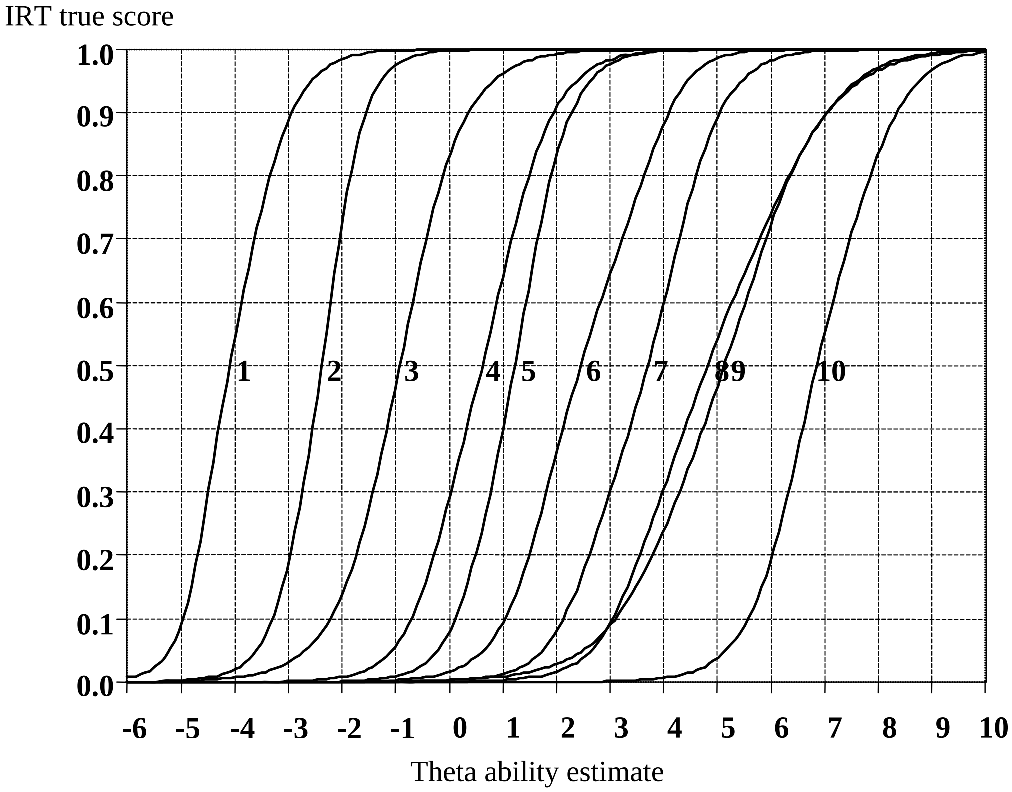
From this analysis, it seems evident that random or internal consistency measurement error was substantially more important than assessor effects on both the BSF-R mental and motor tests used in the ECLS-B. Measurement error in one form or another (random error and assessor effects) accounted for about 42 percent of total variance in initial status on the mental and about 19 percent of the total variance on the motor. The ratios of assessor effects variance to total variance in initial status was about 6 percent for the mental and about 5 percent for the motor. Random or internal consistency error variance represented fully 36 percent of the total variance in initial status on the mental and 14 percent on the motor. For both the BSF-R measures of mental and motor development, internal consistency error proved to be a more important measurement issue than assessor effects in the ECLS-B.

The variance of assessor effects in relation to true-score variance for growth rates was about 11 percent on both the mental and motor tests. On this basis, the impact of assessor effects on growth rates appeared to be moderate in the ECLS-B. However, it should be noted that the potentially important issue of test-retest reliability cannot be addressed in ECLS-B due to the design limitation with data collections at two points in time. In this context, it is perhaps worth noting that test-retest designs are never very reliable since the reliability of growth rates depends fundamentally on, and rises rapidly with, the number of observations obtained for each subject (Bryk and Raudenbush 1987; Willet 1989, 1997). This is particularly unfortunate in a longitudinal study such as the ECLS-B since the impact of existing housing conditions and other contingent social contexts primarily should be assessed in relation to developmental growth rates rather than in relation to developmental status.

## 4.10 BSF-R Proficiency Level Probabilities

One of the convenient features of IRT is that items and persons share the same scale metric. This implies that persons at any given level of ability can be characterized by items at that same threshold. In the ECLS-B, small clusters of items of similar content at roughly the same level of ability were used to represent developmental milestones for young children. Item clusters containing anywhere from 3 to 7 items were identified so that short subscales were built with publisher item calibrations. In the ECLS-B, 10 such subscales were identified for the mental scale and an additional 10 were also identified for the motor scale, as shown in figures 4-25 and 4-26.

Figure 4-25. Response functions for proficiency level subscales representing 10 developmental milestones on the mental scale: 1993

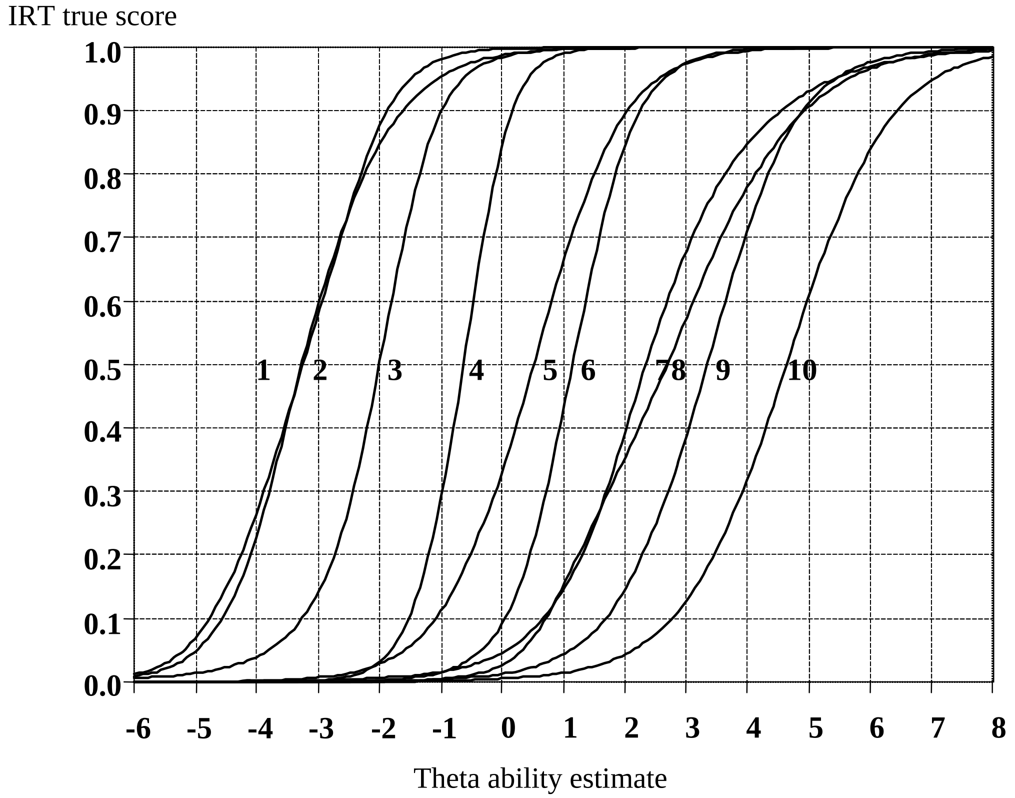


NOTE: Item Response Theory (IRT) item calibrations obtained with unweighted publisher standardization dataset observations. 1 = Explores objects; 2 = Explores purposefully; 3 = Jabbers expressively; 4 = Early problem solving; 5 = Names object; 6 = Receptive vocabulary; 7 = Expressive vocabulary; 8 = Listening/comprehension; 9 = Matching/discrimination; 10 = Early counting/quantitative.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

At higher levels of ability, the test curve characteristic (TCC) for these subscales equaled the total number of items in the scale.[29] To produce proficiency level probabilities, subscale true-scores were divided by the total number of items in the scale. This was equivalent to summing the probabilities computed from each of the subscale component items at a given level of ability and dividing this sum by the total number of items in the subscale, this representing the maximum possible score. This produced a response function rising from zero at low levels of ability to unity at high levels of ability. The response function represented the probability of having reached the developmental milestone represented by the items in the subscale. In this way, a proficiency level subscale performed much like a super-item, and the resulting response function looked much like an item characteristic curve.

[29] Both the ICC and the TCC have response functions that have a similar shape. The only difference is that the ICC represents a probability between 0 and 1, whereas the TCC represents the raw score. Dividing these scores by the maximum possible score on the test, the raw score can be interpreted as a probability. In this sense, proficiency level subscales can behave or act as super-items.

Figure 4-26. Response functions for proficiency level subscales representing 10 developmental milestones on the motor scale: 1993



NOTE: Item Response Theory (IRT) item calibrations obtained with unweighted publisher standardization dataset observations. 1 = Eye-hand coordination; 2 = Sitting; 3 = Pre-walking; 4 = Stands alone; 5 = Skillful walking; 6 = Balance; 7 = Fine motor control; 8 = Uses stairs; 9 = Alternating balance; 10 = Motor planning.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Publisher item calibrations were used to build the proficiency level subscales so that a consistent scale metric was maintained. In the ECLS-B, BSF-R item responses were used to obtain a mean expected a posteriori (EAP) ability estimate. This ability estimate was then applied to the subscale response function to obtain a proficiency level probability. In this fashion, one ability estimate yielded 10 proficiency probabilities, representing the probability that a child had reached each of the 10 developmental milestones. Mean probabilities on both the mental and motor tests are reported in table 4-17.

Table 4-17. Mean proficiency level probabilities for the 10 proficiency level subscales of the BSF-R mental and motor scales at 9 months and 2 years: 2001–02 and 2003–04

| Test | 9 months | 2 years |
|---|---|---|
| Mental scale | | |
| Explores objects | 0.989 | 1.000 |
| Explores purposefully | 0.871 | 1.000 |
| Jabbers expressively | 0.415 | 0.999 |
| Early problem solving | 0.111 | 0.985 |
| Names object | 0.048 | 0.976 |
| Receptive vocabulary | 0.015 | 0.848 |
| Expressive vocabulary | 0.003 | 0.645 |
| Listening/comprehension | 0.001 | 0.373 |
| Matching/discrimination | 0.002 | 0.326 |
| Early counting/quantitative | 0.000 | 0.042 |
| Motor scale | | |
| Eye-hand coordination | 0.914 | 1.000 |
| Sitting | 0.894 | 0.999 |
| Pre-walking | 0.719 | 0.999 |
| Stands alone | 0.326 | 0.998 |
| Skillful walking | 0.182 | 0.928 |
| Balance | 0.092 | 0.897 |
| Fine motor control | 0.046 | 0.563 |
| Uses stairs | 0.036 | 0.489 |
| Alternating balance | 0.015 | 0.310 |
| Motor planning | 0.005 | 0.108 |

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Based on the weighted ECLS-B sample, table 4-17 reports progressively lower probabilities over a sequence of progressively more difficult developmental milestones in cross-section for both the 9-month and 2-year assessments. The one exception to this pattern is from the 9-month probability for listening/comprehension to the 9-month probability for matching/discrimination. For the 9-month data collection, 41.5 percent of the infants babbled, whereas for the 2-year data collection, virtually all of the children surpassed this developmental milestone. Generally speaking, for the 9-month assessment, an appreciably large number of infants performed successfully on the first five subscales on both the mental and motor scales. For the 2-year assessment, appreciably large numbers of children performed satisfactorily on all but the very last of the subscales.

The ECLS-B sample could be used to explicitly model each of these proficiency level probabilities as a function of age. Since probabilities are non-linear, it would be advisable to represent each probability using normal deviates, *probits* or *logits*. Proceeding in this fashion, it would be possible to use regression equations obtained with the weighted ECLS-B sample to search for the mean age at which each developmental level is attained. From a psychometrics perspective, developmental threshold would occur at the age where the predicted probability is 0.5. Another approach would be to consider substantive thresholds where developmental mastery would occur at the age where the predicted probability would be, for example, 0.67 or 0.8.[30] The ECLS-B proficiency level subscales consisting of items taken from the BSID-II, are reported in table 4-18, along with publisher item parameters.

## 4.11 BSF-R Differential Test and Item Functioning

BSF-R 9-month and 2-year item sets were examined for evidence of DTF and DIF in the ECLS-B. This involved comparisons of test performance between a focal group (e.g., African American children) and a reference group (e.g., White children), once individuals in the two groups have been matched or *blocked* on their ability estimates. It was not expected that the different subgroups would perform identically on the same test. Rather, children from two different groups, *who were otherwise identical in terms of their overall ability*, should have had the same probability of obtaining correct responses to the set of items. There should have been no relative advantage or disadvantage in obtaining correct responses based on the child’s subgroup membership.

[30] If the probability of mastery is set very high—for example 0.9 or 0.95—then the mastery age will drift very far away from the age threshold, which is the age where infants are actually acquiring the skill set represented by the developmental milestone. Mastery probabilities of 0.67 or 0.8 are merely suggested as compromises that will keep the mastery age in the vicinity of the age threshold.

Table 4-18. Proficiency level subscales for the BSF-R mental and motor scales: Items in each proficiency level subscale and their IRT ability and discrimination parameters: 2001–02 and 2003–04

| Proficiency level subscale label | BSID-II item number and item label | | Item difficulty *(b)* | Item discrimination *(a)* |
|---|---|---|---|---|
| Explores objects | MEN045 | Picks up Cube | -4.813 | 2.501 |
| | MEN048 | Plays with String | -4.796 | 1.836 |
| | MEN052 | Bangs in Play | -3.930 | 1.158 |
| | MEN053 | Reaches for Second Cube | -3.819 | 1.218 |
| | MEN055 | Lifts Inverted Cup | -4.361 | 1.402 |
| | MEN057 | Picks up Cube Deftly | -3.773 | 1.167 |
| | MEN059 | Manipulates Bell, Showing Interest in Detail | -2.963 | 1.643 |
| Explores purposefully | MEN062 | Pulls String Adaptively to Secure Ring | -2.652 | 1.096 |
| | MEN065 | Retains Two of Three Cubes for 3 Seconds | -2.405 | 1.616 |
| | MEN066 | Rings Bell Purposely | -2.393 | 1.546 |
| | MEN069 | Looks at Pictures in Book | -2.192 | 1.805 |
| Jabbers expressively | MEN076 | Jabbers Expressively | -0.749 | 0.940 |
| | MEN078 | Vocalizes Four Different Vowel-Consonant Combinations | -1.114 | 0.838 |
| | MEN081 | Responds to Spoken Request | -1.015 | 1.233 |
| Early problem solving | MEN089 | Puts Six Beads in Box | -0.280 | 1.521 |
| | MEN095 | Puts Nine Cubes in Cup | 0.692 | 0.953 |
| | MEN102 | Retrieves toy (Visible Displacements) | 1.021 | 1.099 |
| | MEN104 | Uses Rod to Attain Toy | 1.012 | 1.177 |
| Names object | MEN100 | Uses Two Different Words Appropriately | 0.734 | 1.316 |
| | MEN101 | Shows Shoes, Other Clothing, or Object | 0.746 | 1.569 |
| | MEN106 | Uses Word(s) to Make Wants Known | 1.613 | 1.969 |
| | MEN110 | Names One Object | 1.732 | 1.186 |
| Receptive vocabulary | MEN099 | Points to Two Pictures | 1.944 | 1.066 |
| | MEN108 | Points to Three of Doll’s Body Parts | 1.919 | 1.228 |
| | MEN122 | Points to Five Pictures | 3.660 | 1.452 |

See note at end of table.

Table 4-18. Proficiency level subscales for the BSF-R mental and motor scales: Items in each proficiency level subscale and their IRT ability and discrimination parameters: 2001–02 and 2003–04—Continued

| Proficiency level subscale label | BSID-II item number and item label | | Item difficulty *(b)* | Item discrimination *(a)* |
|---|---|---|---|---|
| Expressive vocabulary | MEN111 | Combines Word and Gesture | 2.487 | 1.615 |
| | MEN114 | Uses A Two-Word Utterance | 3.109 | 0.896 |
| | MEN121 | Uses Pronoun(s) | 3.975 | 1.197 |
| | MEN126 | Names Three Objects | 4.144 | 1.283 |
| | MEN133 | Names Five Pictures | 4.365 | 1.160 |
| Listening/comprehension | MEN131 | Attends to Story | 3.511 | 1.397 |
| | MEN134 | Displays Verbal Comprehension | 4.459 | 0.939 |
| | MEN140 | Understands Two Prepositions | 5.184 | 0.865 |
| | MEN142 | Multiple-Word Utterances Response to Picture Book | 6.524 | 0.932 |
| Matching/discrimination | MEN125 | Matches Pictures | 3.967 | 1.003 |
| | MEN128 | Matches Three Colors | 4.052 | 0.667 |
| | MEN137 | Matches Four Colors | 5.194 | 0.801 |
| | MEN144 | Discriminates Pictures I | 5.593 | 1.293 |
| | MEN151 | Discriminates Pictures II | 6.521 | 0.794 |
| Early counting/quantitative | MEN141 | Understands Concept of One | 5.925 | 1.327 |
| | MEN146 | Counts (Number Names) | 6.464 | 1.871 |
| | MEN147 | Compares Masses | 6.486 | 1.105 |
| | MEN152 | Repeats Three Number Sequences | 6.770 | 0.857 |
| | MEN156 | Understands Concept of More | 7.520 | 1.193 |
| | MEN159 | Counts (Stable Number order) | 6.933 | 1.276 |
| | MEN164 | Counts (Cardinality) | 8.039 | 1.364 |
| Eye-hand coordination | MOT031 | Uses Partial Thumb Opposition to Grasp Cube | -3.845 | 1.267 |
| | MOT032 | Attempts to Secure Pellet | -3.741 | 1.135 |
| | MOT041 | Uses Whole Hand to Grasp Pellet | -2.859 | 0.919 |
| | MOT049 | Uses Partial Thumb Opposition to Grasp Pellet | -2.642 | 0.720 |

See note at end of table.

Table 4-18. Proficiency level subscales for the BSF-R mental and motor scales: Items in each proficiency level subscale and their IRT ability and discrimination parameters: 2001–02 and 2003–04—Continued

| Proficiency level subscale label | BSID-II item number and item label | | Item difficulty *(b)* | Item discrimination *(a)* |
|---|---|---|---|---|
| Sitting | MOT022 | Sits with Slight Support for 10 Seconds | -4.469 | 1.170 |
| | MOT028 | Sits Alone Momentarily | -4.142 | 1.162 |
| | MOT034 | Sits Alone for 30 Seconds | -3.195 | 1.082 |
| | MOT036 | Sits Alone Steadily | -3.260 | 0.974 |
| | MOT043 | Moves Forward Using Prewalking Methods | -2.883 | 1.023 |
| | MOT051 | Moves from Sitting to Creeping Position | -2.400 | 1.520 |
| Pre-walking | MOT044 | Supports Weight Momentarily | -2.751 | 0.625 |
| | MOT045 | Pulls to Standing Position | -2.488 | 1.219 |
| | MOT046 | Shifts Weight while Standing | -2.257 | 1.403 |
| | MOT052 | Raises Self to Standing Position | -2.001 | 1.895 |
| | MOT053 | Attempts to Walk | -1.724 | 1.223 |
| | MOT054 | Walks Sideways while Holding on to Furniture | -1.604 | 1.801 |
| Stands alone | MOT059 | Stands up I | -0.490 | 1.395 |
| | MOT060 | Walks with Help | -1.186 | 1.676 |
| | MOT061 | Stands Alone | -0.669 | 1.734 |
| | MOT062 | Walks Alone | -0.344 | 2.274 |
| Skillful walking | MOT063 | Walks Alone with Good Coordination | -0.295 | 0.917 |
| | MOT067 | Walks Backward | 0.760 | 1.088 |
| | MOT071 | Walks Sideways | 1.020 | 0.829 |
| Balance | MOT065 | Squats Briefly | 1.386 | 1.087 |
| | MOT068 | Stands up II | 0.993 | 1.070 |
| | MOT072 | Stands on Right Foot with Help | 1.006 | 1.267 |
| | MOT073 | Stands on Left Foot with Help | 1.213 | 1.371 |
| Fine motor control | MOT074 | Uses Pads of Fingertips to Grasp Pencil | 2.037 | 1.077 |
| | MOT075 | Uses Hand to Hold Paper in Place | 2.199 | 0.763 |
| | MOT090 | Grasps Pencil at Nearest End | 3.227 | 0.513 |

See note at end of table.

Table 4-18. Proficiency level subscales for the BSF-R mental and motor scales: Items in each proficiency level subscale and their IRT ability and discrimination parameters: 2001–02 and 2003–04—Continued

| Proficiency level subscale label | BSID-II item number and item label | | Item difficulty *(b)* | Item discrimination *(a)* |
|---|---|---|---|---|
| Uses stairs | MOT069 | Walks down Stairs with Help | 1.058 | 1.420 |
| | MOT079 | Walks up Stairs Alone, Placing Both Feet on Each Step | 2.632 | 0.949 |
| | MOT080 | Walks down Stairs Alone, Placing Both Feet on Step | 3.104 | 0.867 |
| | MOT095 | Walks up Stairs, Alternating Feet | 4.210 | 0.675 |
| Alternating balance | MOT082 | Stands Alone on Right Foot | 2.887 | 0.950 |
| | MOT083 | Stands Alone on Left Foot | 2.959 | 0.754 |
| | MOT086 | Swings Leg to Kick Ball | 3.803 | 1.194 |
| | MOT089 | Walks on Tiptoe for Four Steps | 3.584 | 0.697 |
| Motor planning | MOT088 | Laces Three Beads | 3.892 | 0.643 |
| | MOT091 | Imitates Hand Movements | 4.208 | 0.830 |
| | MOT093 | Manipulates Pencil in Hand | 4.524 | 0.866 |
| | MOT096 | Copies Circle | 4.506 | 0.694 |
| | MOT098 | Imitates Postures | 5.059 | 0.820 |
| | MOT101 | Buttons One Button | 5.382 | 0.787 |

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

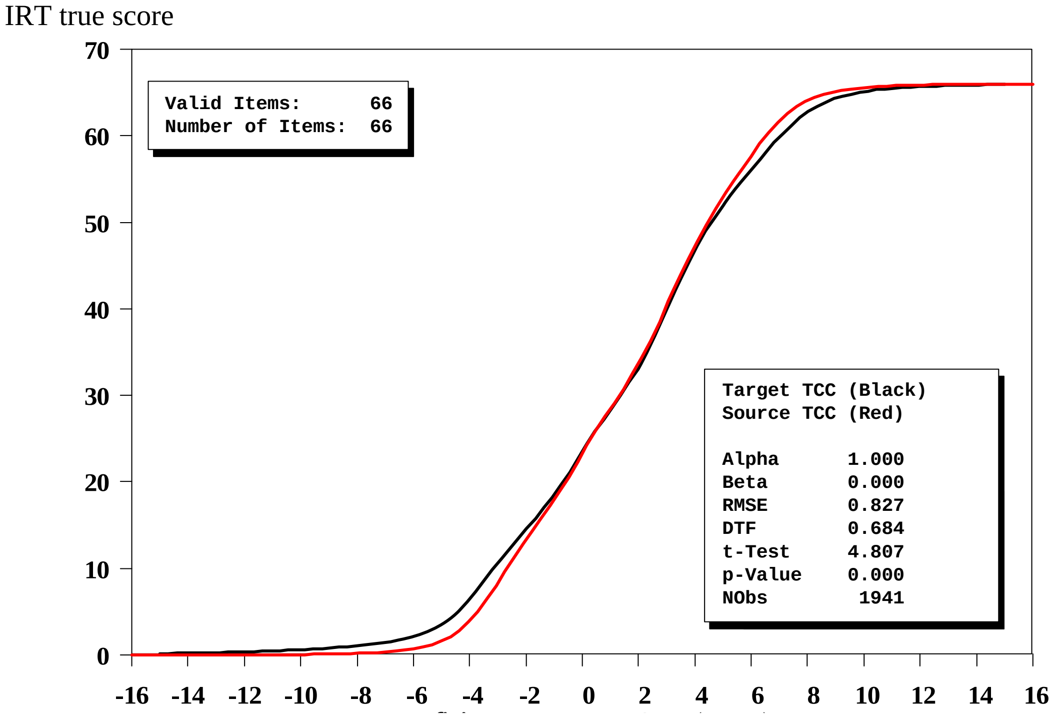
A test is said to exhibit DTF when individuals having the same ability, but from different groups, fail to obtain the same number of correct responses. IRT provides a unified framework for investigating issues of statistical bias at both the test and item levels. A test shows evidence of statistical bias when, at the same level of ability, two groups fail to obtain the same score. DTF is examined in the ECLS-B using parametric IRT procedures developed by Raju, van der Linden, and Fleer (1995).

For this purpose, a series of separate response vector files were created for focal minority groups and reference majority groups using observations obtained at 9 months and 2 years. Each file was then scored separately using identical sets of BSF-R item parameters. The scoring effectively classifies each observation by ability level. As each observation was scored separately in each group, marginal likelihoods were accumulated for each item response across all levels of ability. Once all observations were scored in this fashion, new sets of IRT parameters were fitted to the marginal likelihoods in a single iteration. The new sets of item parameters represented the response characteristics for each respective focal or reference group across all levels of ability.

The issue to be addressed in DTF analysis was whether children at the same level of ability on average obtained the same number-right score on the same test. This issue was examined in IRT by comparing the TCCs for the two groups. The TCC is the sum of the ordinates of the ICCs at each level of ability, $\xi = \sum_{j=1}^{n} P_j(\theta)$. The TCC represents the expected number of correct responses, expressed in raw score metric, equivalent to the number of items that would be answered correctly on a test. Any misalignment of TCCs reveals evidence of DTF. The total number-right score at each level of ability was examined by comparing IRT true-scores for each focal and reference group comparison.

The new sets of item parameter estimates were used for these group comparisons. The TCC for the focal (source) and reference (target) tests were compared across all levels of ability. The weighted sum of squared differences between the source and target test characteristic curves was used as a DTF index. The DTF coefficient quantified the degree of misalignment between the two curves, expressed in squared raw score units. The square root of the DTF coefficient was an RMSE, expressed in raw score units. The magnitude of RMSE values were interpreted bearing in mind the maximum raw score possible or the average raw score on the test in question. These residual measures of dispersion around the target TCC were the DTF statistics most frequently reported in the literature, (Raju, van der Linden, and Fleer 1995) as shown in figure 4-27.

Figure 4-27. DTF analysis, showing mental test characteristic curves for BSF-R Asian focal group and White reference group before equating: 2001–02 and 2003–04



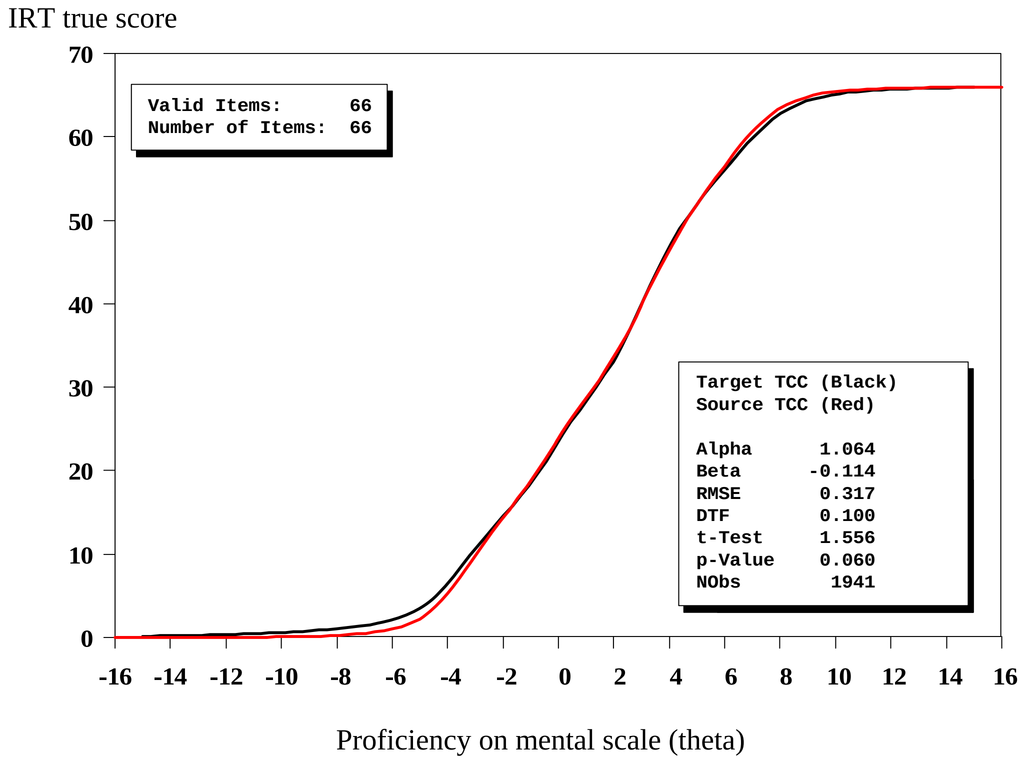
Proficiency on mental scale (theta)

NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

However, from a practical point of view, researchers were more concerned about the overall magnitude and direction of statistical bias as this may have affected ability estimates. With large samples such as the ECLS-B, virtually any DTF coefficient was statistically significant. This implied that it was appropriate to generalize from the focal and reference group samples to the same groups in the ECLS-B population and affirmed that *at least some DTF greater than zero* exists when these instruments were used with these subgroups of the population. If some statistical bias existed in the population, then it was often more meaningful to ascertain the overall direction and magnitude of this statistical bias.

Thus, it was also helpful to consider the average overall difference between test scores in the two groups in terms of the population standard deviation units expressed by the IRT scale metric. Estimates of the average overall statistical bias were obtained with IRT true-score equating, which showed the linear transformation of origin and scale that would be needed to align the source (focal) and target (reference) tests. In the context of DTF analysis, equating constants α (slope) and β (origin) were expressions of the overall statistical bias expected when the assessment instrument was used with the focal group. The overall group effect was represented by the intercept coefficient β, whereas a group by ability interaction effect was represented by the slope coefficient α, as shown in figure 4-28.

Figure 4-28. DTF analysis, showing mental test characteristic curves for BSF-R Asian focal group and White reference group after equating: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Both equating coefficients α and β were reported in population standard deviation units and were, thus, effect-size measures of the average statistical bias of focal group ability estimates relative to reference group ability estimates. Under conditions of perfect test alignment in the two groups and no evidence of DTF, the expectation was to find origin β = 0 and slope α = 1, indicating that no statistical bias was present when comparing the two groups. When β ≠ 0, a group effect was present, demonstrating that some spurious group-related trait unrelated to the trait purportedly measured by the test was also captured by the assessment instrument. When $\alpha \neq 1$, a group-ability interaction effect unrelated to the test objective was also present. Although these coefficients were reported less frequently in the DTF literature, conceptually they were very useful and easy to understand.

DTF statistics for the ECLS-B are reported in table 4-19 for the three focal and reference group comparisons considered in the ECLS-B. With the large sample size available in the ECLS-B, many DTF and RMSE measures were statistically significant prior to equating, whereas virtually no measure of dispersion between TCCs was statistically significant once $\alpha$ and $\beta$ were used to relate the focal and reference groups. This demonstrates rather conclusively that a group effect and group by ability interaction effect accounted for virtually all of a distinctly linear form of statistical bias. At risk of little or no simplification, the statistical bias represented by the $\beta$ coefficient, conveniently expressed in population standard deviation units, was generally sufficient to summarize the overall difference in test performance.

Inspecting the $\beta$ coefficient values in table 4-19 reveals statistical biases that ranged from minute to small for the nine focal and reference group comparisons considered in this exercise. These comparisons included three race-ethnicity focal groups (African Americans, Hispanics, and Asians), one comparison each for gender, premature and SES focal groups, and three comparisons for maternal-child attachment behavior focal groups as measured and identified by the TAS-45 in the ECLS-B. Small statistical biases affected each minority group on the mental test, ranging from -0.010 population standard deviation bias for the low-SES group down to -0.221 for attachment style D (disorganized) children. Indeed, there was evidence of a modest amount of DIF for both the C and D attachment styles on both ECLS-B instruments. By contrast, there was little, if any, evidence of statistical bias of appreciable magnitude for African Americans or any other minority ethnic group on either ECLS-B test, except possibly a small -0.114 population standard deviation bias for Asians on the mental test.

Table 4-19. BSF-R differential test functioning (DTF) statistics (DTF index, RMSE, Alpha, and Beta) for focal group—reference group comparisons on the mental and motor scales: 2001–02 and 2003–04

| Focal—reference group comparison | Statistic | Mental scale | Motor scale |
|---|---|---|---|
| Black — White | | | |
| | DTF | 0.080 | 0.057 |
| | RMSE | 0.282 | 0.239 |
| | Alpha | 1.004 | 1.023 |
| | Beta | -0.058 | -0.025 |
| Hispanic — White | | | |
| | DTF | 0.055 | 0.005 |
| | RMSE | 0.235 | 0.073 |
| | Alpha | 0.994 | 1.001 |
| | Beta | -0.019 | 0.005 |
| Asian — White | | | |
| | DTF | 0.684 | 0.123 |
| | RMSE | 0.827 | 0.351 |
| | Alpha | 1.064 | 1.020 |
| | Beta | -0.114 | -0.049 |
| Female — Male | | | |
| | DTF | 0.040 | 0.469 |
| | RMSE | 0.199 | 0.685 |
| | Alpha | 0.992 | 1.006 |
| | Beta | 0.042 | 0.140 |
| Low SES — High SES | | | |
| | DTF | 0.024 | 0.041 |
| | RMSE | 0.156 | 0.202 |
| | Alpha | 0.997 | 1.014 |
| | Beta | -0.010 | -0.034 |
| Premature — Full Term | | | |
| | DTF | 0.132 | 0.168 |
| | RMSE | 0.363 | 0.409 |
| | Alpha | 1.022 | 1.020 |
| | Beta | -0.025 | -0.051 |
| A — B Attachment Style | | | |
| | DTF | 0.140 | 0.064 |
| | RMSE | 0.375 | 0.253 |
| | Alpha | 1.002 | 1.008 |
| | Beta | -0.072 | -0.049 |
| C — B Attachment Style | | | |
| | DTF | 1.058 | 0.379 |
| | RMSE | 1.029 | 0.616 |
| | Alpha | 1.015 | 1.010 |
| | Beta | -0.216 | -0.111 |

See note at end of table.

Table 4-19. BSF-R differential test functioning (DTF) statistics (DTF index, RMSE, Alpha, and Beta) for focal group—reference group comparisons on the mental and motor scales: 2001–02 and 2003–04—Continued

| Focal—reference comparison | Statistic | Mental scale | Motor scale |
|---|---|---|---|
| D — B Attachment Style | | | |
| | DTF | 1.082 | 0.472 |
| | RMSE | 1.040 | 0.687 |
| | Alpha | 1.016 | 0.990 |
| | Beta | -0.221 | -0.127 |

NOTE: An item exhibits DIF "if individuals of the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, Swaminathan, and Rogers 1991, p. 110). Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

DTF and RMSE coefficients reported in table 4-19 are primarily reflections of this same, relatively small statistical bias, as this affected the number-right raw score. The largest RMSE value was a 1.040 raw score point disparity at any level of ability for the D—B attachment style comparison. These coefficients revealed nothing about the direction of statistical bias, whereas coefficient β showed that this was negative or zero for virtually all groups except possibly for females on the motor test. Several DTF and RMSE coefficients were statistically significant prior to a linear transformation, whereas rarely were any of the differences statistically significant after a linear transformation of origin and of scale. This showed that the principal difference between focal and reference groups was usually a question of systematic linear statistical bias. However, it is important to emphasize that DTF should be observed and quantified prior to any such linear transformation.

DIF has also been examined in the ECLS-B. DIF identified individual items that showed an unexpectedly large difference in the probability of a correct response when comparing individuals in the focal and reference groups at the same level of ability. For the ECLS-B sample, DIF indices were calculated using the parametric IRT procedures developed by Raju, van der Linden, and Fleer (1995). Table 4-20 summarizes these results, showing all of the items on the BSF-R that exhibited weighted RMSEs of 0.10 or more. Items showing lower levels of DIF have been excluded from the table in order to save space. Like the DTF indices presented previously, these weighted root mean square NC-DIF indices reflect the magnitude of the distance between ICCs but not the direction of bias.

Table 4-20. BSF-R differential item functioning (DIF) for mental and motor items that exhibited weighted root mean squared errors (RMSEs) of 0.10 or more: 2001–02 and 2003–04

| | | Selected demographic characteristics | | | | | | Attachment style classifications | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Item | Label | Premature | Low SES | Female | Black | Hispanic | Asian | A_ Avoidant | C_ Ambivalent | D_ Disorganized |
| MEN058 | Retains Two Cubes for 3 Seconds (Basal) | † | † | † | † | † | † | 0.112 | † | † |
| MEN099 | Points to Two Pictures (Basal) | † | † | † | † | † | † | † | 0.117 | 0.134 |
| MEN101 | Shows Shoes, Other Clothing, or Object (Ceiling) | † | † | † | † | † | † | † | 0.170 | † |
| MEN102 | Retrieves toy (Visible Displacements) (Ceiling) | † | † | † | † | † | 0.104 | † | † | † |
| MEN104 | Uses Rod to Attain Toy (Ceiling) | † | † | † | † | † | † | † | 0.106 | † |
| MEN107 | Follows Directions (Doll) (Basal) | † | † | † | † | † | † | † | † | 0.116 |
| MEN108 | Points to Three of Dolls Body Parts (Basal) | † | † | † | † | † | † | † | 0.113 | 0.139 |
| *MEN110* | *Names One Object (Basal)* | *†* | *†* | *†* | *†* | *0.105* | *†* | *†* | *†* | *†* |
| MEN117 | Imitates A Two-Word Sentence (Core) | † | † | † | 0.131 | † | † | † | † | † |
| MEN122 | Points to Five Pictures (Core) | † | † | † | | † | † | † | † | 0.107 |
| MEN123 | Builds Tower of Six Cubes (Core) | 0.131 | † | † | † | 0.116 | † | † | 0.106 | 0.140 |
| MEN124 | Discriminates Book, Cube and Key (Core) | † | † | † | † | † | † | † | † | 0.131 |
| MEN128 | Matches Three Colors (Core) | † | † | † | 0.132 | † | † | † | † | † |
| MEN131 | Attends to Story (Core) | † | † | † | † | † | † | 0.141 | 0.140 | 0.190 |
| MEN137 | Matches Four Colors (Core) | † | † | † | 0.140 | † | † | † | † | † |
| MEN141 | Understands Concept of One (Core) | † | † | † | † | 0.102 | 0.105 | † | † | † |
| MOT074 | Uses Pads of Fingertips to Grasp Pencil (Core) | † | † | 0.137 | † | 0.117 | † | † | † | † |
| MOT075 | Uses Hand to Hold Paper in Place (Core) | † | 0.109 | 0.137 | † | † | † | † | 0.157 | 0.141 |
| MOT084 | Walks Forward on Line (Core) | † | † | † | † | † | † | † | 0.107 | 0.116 |
| MOT090 | Grasps Pencil at Nearest End (Core) | † | † | 0.144 | † | † | † | † | † | † |
| MOT093 | Manipulates Pencil in Hand (Core) | † | † | 0.132 | † | † | † | † | † | † |
| MOT102 | Stands Alone on Left Foot for 4 Seconds (Ceiling) | † | † | † | † | † | † | † | 0.211 | † |

† Not applicable.

NOTE: An item exhibits DIF “if individuals of the same ability, but from different groups, do not have the same probability of getting the item right” (Hambleton, Swaminathan, and Rogers 1991, p. 110). Items are considered to exhibit expressive DIF when the weighted root mean squared difference between focal and reference group item characteristic curves exceeds 0.10 or 10 percentage points. SES = socioeconomic status.

SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

*This page is intentionally left blank.*

# 5. BAYLEY SHORT FORM–RESEARCH EDITION SCORES

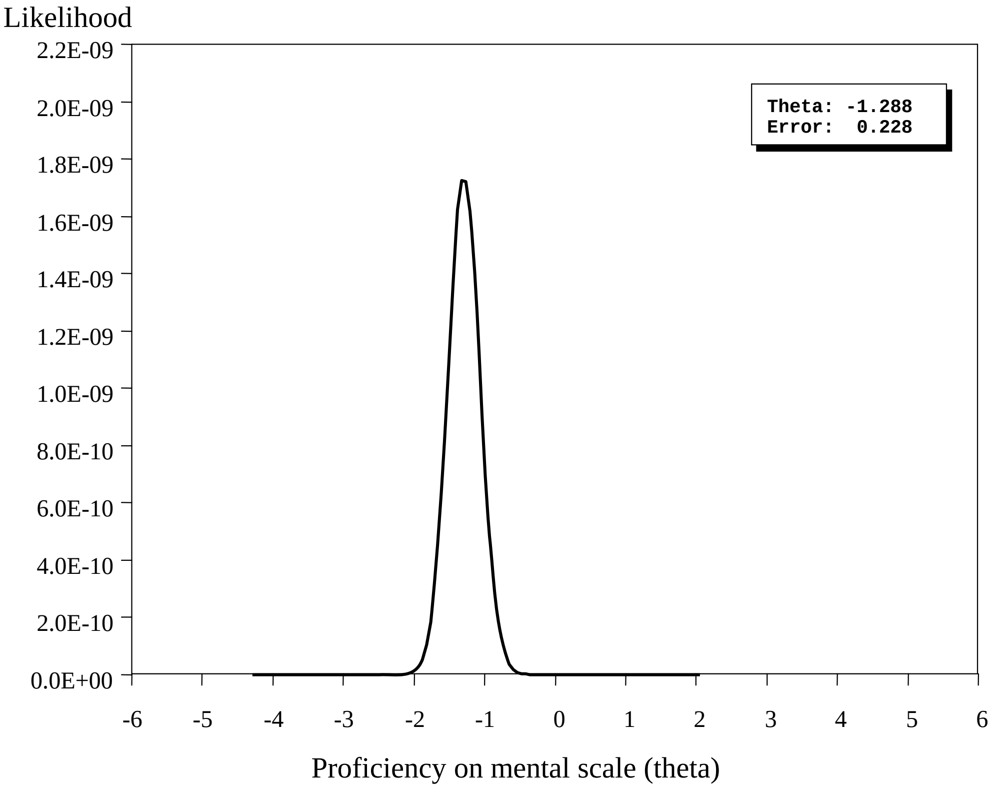
The Bayley Scales of Infant Development, Second Edition (BSID-II), published by The Psychological Corporation, is an assessment of developmental status and not an intelligence (IQ) test in which a single total score is obtained that represents an individual's verbal and performance intelligence. There is no total score on the BSID-II. Rather, there are separate scores for the mental scale (Mental Development Index) and for the motor scale (Psychomotor Development Index). Similarly, there are separate scores for the Bayley Short Form–Research Edition (BSF-R) mental scale and motor scale, but there is no single BSF-R total score. This conforms to standard scoring procedures for the BSID-II. The BSF-R scores on the longitudinal 9-month–2-year data file are summarized in the following sections.

## 5.1 BSF-R Scoring and Ability Estimates

In Item Response Theory (IRT), the item characteristic curve (ICC) represents the probability of a correct response, P($x$ = 1), across all levels of ability. Item calibrations model the probabilities of a correct response on each of several items. In probability theory, for any two independent events *A* and *B*, the probability of both events occurring simultaneously is given by the product of the probability of either event occurring separately: P (*A* & *B*) = P(*A*)P(*B*). In IRT, it is similarly assumed that item responses are independent events. In other words, the answer to any one item provides no information that can be used by the examinee to answer any other item.

In the fashion of independent events *A* and *B*, the likelihood of a set of responses is obtained by multiplying all of the corresponding item probabilities in series. If the examinee gets the item right, then the 2-parameter logistic (2-PL) function estimate of P($x$ = 1) is used. If not, the IRT estimate of P($x$ = 0) = 1 − P($x$ = 1) is used. Since the logistic function is a continuous function, the likelihood of any response vector can be estimated across all ability levels. The new distribution is known as the response likelihood distribution. An example of a likelihood distribution for a child in the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) is shown in figure 5-1, which shows that the likelihood of a response vector is quite small at any level of ability but appreciably smaller at some levels than at others. Moreover, when the items and response vectors are informative (that is when they contain information useful for determining ability level), the range of more prominent likelihood values is constrained within a relatively short range. When the likelihood distribution is sharply concentrated, its graphical representation is similar to a spike. In this particular example, the child is most likely to be found in the lower tail of the ability distribution to the left of the figure. The largest likelihood would provide a good guess of this child's ability, and indeed the maximum likelihood is often used as if it were *the* ability estimate for a given observation. On the basis of maximum likelihood, the ability level of the child represented by the figure would be $\overline{\theta}_i$ = -1.288.

Figure 5-1. Response likelihood function for a specific examinee on the BSF-R mental scale: 2003–04



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.

On the other hand, it can be seen that the likelihood around this central tendency forms its own distribution. Indeed, by calculating the standard deviation of the likelihood distribution, the standard error of measurement is obtained, which is reported to be SE($\overline{\theta}_i$) = 0.228 in the figure. A typical error, that is, the average error, can be expected to lie roughly within a third of a population standard deviation to either side of the maximum likelihood.

## 5.2 Expected a Posteriori Ability Estimate

The expected a posteriori (EAP) estimate of ability for an individual *i* is

$$\bar{\theta}_i \cong \frac{\sum_{k=1}^{q} X_k P(\mathbf{x}_i \mid X_k) A(X_k)}{\sum_{k=1}^{q} P(\mathbf{x}_i \mid X_k) A(X_k)},$$

where $P(\mathbf{x}_i \mid X_k)$ represents the likelihood of response vector $\mathbf{x}_i$ at point $X_k$ on the ability axis.[1] This is also known as the Bayes estimate of the posterior distribution of $\theta$, given response pattern $\mathbf{x}_i$. The EAP estimate is approximated using Gaussian quadrature, where $A(X_k)$ are normal ordinate weights for *q* points $X_k$ spanning the ability distribution for the age group containing member *i*.

## 5.3 Expected a Posteriori Standard Error of Measurement

The error variance of the EAP ability estimate is

$$\sigma_e^2 \cong \frac{\sum_{k=1}^{q} (X_k - \bar{\theta}_i)^2 P(\mathbf{x}_i \mid X_k) A(X_k)}{\sum_{k=1}^{q} P(\mathbf{x}_i \mid X_k) A(X_k)}.$$

The standard error of measurement for EAP ability estimates is the square root of this value. The standard error represents the measurement error but ignores errors that may result from equating to publisher scale metric.

[1] This formula provides an operational definition showing how EAP ability estimates were calculated with quadrature points. For a conceptual discussion of EAP ability estimates, please see the footnote on p. 4-23.

## 5.4 BSF-R Scale Scores and *T* Scores

The BSF-R scores provided in the longitudinal 9-month–2-year data file have been recalibrated from the 9-month cross-sectional file. These longitudinal (9-month to 2-year) recalibrated scores may have changed for an individual child, but relative standing should not have changed (or, if so, only minimally). The analyst is encouraged to use these recalibrated scores exclusively because they supersede the 9-month cross-sectional scores.

The Psychological Corporation uses number-right scoring for the BSID-II mental and motor scales. Raw scores are calculated by adding the number of the item immediately prior to the first item in the administered item set to the total number of correct responses in each administered item set. In essence, the child is automatically given credit for all items from the younger (i.e., easier) age sets. For example, if the BSID-II mental scale were administered beginning with item number 63, and the child was able to complete 6 items correctly within that age set, then the child would receive a raw score of 62 + 6 = 68 points.

To compare the development levels of children of different ages, The Psychological Corporation provides development index numbers that have a mean of 100 and a standard deviation of 15 in each age group. Development index numbers are obtained in BSID-II by using the raw score to find the corresponding development index number in a lookup table provided in The Psychological Corporation documentation. The child's age in years, months, and days is used to determine which page of the table should be used.

In the ECLS-B, IRT true scores (called scale scores in the data file) substitute for BSID-II raw scores. For each EAP ability estimate $\overline{\theta}_i$, obtained with the BSF-R, a corresponding IRT true score $\xi_i$ is calculated by summing the expected probability of a correct response $\xi_i = \sum_{j=1}^{n} P_j(\overline{\theta}_i)$ for all items $j$ = 1 .. $n$ comprising the publisher scale. The number-right true score $\xi_i$ is then used to assign a corresponding development index number. In the ECLS-B, a parametric model based on The Psychological Corporation documentation is used for this purpose, instead of a lookup table. The development indices (called *T* scores) provided in the ECLS-B data file have a mean of 50 and a standard deviation of 10 and should be regarded as approximate values due to any errors associated with $\overline{\theta}_i$. The *T* scores provide a convenient means to examine the developmental levels of children of different ages, equivalent to the developmental index scores provided with BSID-II.

## 5.5 ECLS-B Proficiency Level Probabilities

The BSF-R item response models provide interval scales along which every item and every child is positioned. The substantive significance of EAP ability estimates $\overline{\theta}_i$ can be determined by examining the task content of items positioned at the same level of difficulty. Item clusters, representing tasks positioned at the same or similar levels of ability, are examined in this way for evidence of a common pattern of behavior.

To the extent that a consistent interpretation of the items is possible, item clusters can be used to represent specific levels of proficiency. These proficiency levels become benchmark performance standards or anchor points used to interpret scale values and give them a specific behavioral significance. They provide EAP ability estimates $\overline{\theta}_i$ with a tangible, real-world reference. The identification of proficiency levels often helps to establish a scale as a medium of exchange so that measurement results can be easily comprehended and communicated.

The BSF-R has been developed to provide practical measures of children's mental and motor development and to reproduce as closely as possible measures obtained with the BSID-II. Item clusters have been selected from the BSID-II to help interpret EAP ability estimates at specific levels of proficiency. BSID-II proficiency probabilities have been created by selecting item subsets from the BSID-II mental and motor development scales to form item clusters. That is, the proficiency probabilities, though based on the BSF-R calibration, were formed using the full complement of BSID-II items. This means that the item itself did not necessarily have to be administered as part of the BSF-R to be included in the proficiency probabilities. This is possible due to the use of IRT modeling (exhibits 3-1 and 3-2 show which items were part of the BSF-R). Theoretical considerations, item content, and item difficulty parameters were used to select item subsets that would be as internally consistent as possible. Similar considerations were invoked to attribute a behaviorally significant name for each item cluster.

The analyst is also cautioned that the availability of longitudinal data and the recalibration of the 9-month cross-sectional scores have led to changes in the items comprising some of the mental and motor proficiency levels that were provided on the 9-month data file. The 9-month cross-sectional scores in the previous release were the most accurate that were available at that time. However, with the addition of longitudinal data, the new proficiency levels are preferred. The new 9-month proficiencies supercede those issued in the previous 9-month data file. Exhibit 5-1 lists the items that compose the proficiency probabilities included on both the cross-sectional 9-month and the longitudinal 9-month–2-year data files and notes where changes were made to the contents of specific probabilities.

After consideration to all these issues, the proficiency level probability scales were identified and are presented in table 5-1, which also includes the 9-month and 2-year means.

Subscales for the ECLS-B were constructed using publisher item calibrations by selecting the appropriate subsets of items. By using the publisher item calibrations, the subscale score metric remains identical to that used in the corresponding publisher main scale. Subscales vary in length from three to seven items, depending on the availability of suitable items in BSID-II. The item clusters can be used to calculate subscale true scores, information functions, and standard errors of measurement as with any IRT scale. However, the purpose of the subscales is to define proficiency level probabilities.

A performance level can be defined at a point on the ability scale where two-thirds of the items in the subscale are expected to be answered correctly. This is the point where the IRT true score reaches 67 percent of the total number of items included in the subscale. For example, for a subscale with four items, the performance level is defined at the point on the ability scale where the IRT true score reaches $0.67 \times 4 = 2.66$ correct responses. When 67 percent of the items are expected to be answered correctly, most of the tasks will be completed successfully, and it can be said that mastery of this performance level has been achieved.

The selection of performance level subscales is limited by the availability of items in the corresponding published mental and motor scales. For this reason, it is not possible to define performance milestones at equal scale intervals. As shown in table 5-1, in the case of mental performance, Receptive vocabulary represents a low level of development for 2-year-olds (because almost 85 percent of 2-year-olds can do this). For all practical intents and purposes, Explores objects can be used to identify children with deficient development (because virtually all children can do this by 2 years). At the other extreme, Matching/discrimination and Early counting/quantitative identify children who are highly developed (because relatively fewer can do them). This leaves Expressive vocabulary and Listening/comprehension as milestone events that are more appropriate for 2-year-olds.

Exhibit 5-1. Changes to 9-month proficiency levels following recalibration using longitudinal 9-month–2-year dataset, original 9-month proficiency level variable names, new 9-month proficiency level variable names, and 2-year proficiency level variable names included within each proficiency: 2001–02 and 2003–04

| Original 9-month proficiencies in the 9-month data file | 9-month recalibrated proficiencies in the longitudinal 9-month–2-year data file | 2-year proficiencies in the longitudinal 9-month–2-year data file |
|---|---|---|
| Mental scale | | |
| **X1MTL1 Explores objects** | **X1MTL_A Explores objects** | **X2MTL_A Explores objects** |
| MEN045 Picks up cubes | MEN045 Picks up cubes | MEN045 Picks up cubes |
| MEN048 Plays with string | MEN048 Plays with string | MEN048 Plays with string |
| MEN055 Lifts inverted cup | MEN055 Lifts inverted cup | MEN055 Lifts inverted cup |
| MEN057 Picks up cube deftly | MEN057 Picks up cube deftly | MEN057 Picks up cube deftly |
| MEN053 Reaches for 2nd cube | MEN053 Reaches for 2nd cube | MEN053 Reaches for 2nd cube |
| MEN052 Bangs in play | MEN052 Bangs in play | MEN052 Bangs in play |
| **X1MTL2 Explores purposefully** | **X1MTL_B Explores purposefully** | **X2MTL_B Explores purposefully** |
| MEN059 Manipulates bell | MEN059 Manipulates bell | MEN059 Manipulates bell |
| MEN062 Pulls string adaptively | MEN062 Pulls string adaptively | MEN062 Pulls string adaptively |
| MEN065 Retains 2 of 3 cubes | MEN065 Retains 2 of 3 cubes | MEN065 Retains 2 of 3 cubes |
| MEN066 Rings bell purposefully | MEN066 Rings bell purposefully | MEN066 Rings bell purposefully |
| MEN069 Looks at pictures in book | MEN069 Looks at pictures in book | MEN069 Looks at pictures in book |
| **X1MTL3 Babbles** | **X1MTL_C Jabbers Expressively** | **X2MTL_C Jabbers Expressively** |
| MEN061 Vocalizes 3 vowels | [MEN061 deleted] | [MEN061 deleted] |
| MEN078 Vocalizes 4 vowel/ consonant combinations | MEN078 Vocalizes 4 vowel/ consonant combinations | MEN078 Vocalizes 4 vowel/ consonant combinations |
| MEN081 Responds to request | MEN081 Responds to request | MEN081 Responds to request |
| MEN076 Jabbers expressively | MEN076 Jabbers expressively | MEN076 Jabbers expressively |
| **X1MTL4 Early problem solving** | **X1MTL_D Early problem solving** | **X2MTL_D Early problem solving** |
| MEN089 Puts 6 beads in box | MEN089 Puts 6 beads in box | MEN089 Puts 6 beads in box |
| MEN095 Puts 9 cubes in cup | MEN095 Puts 9 cubes in cup | MEN095 Puts 9 cubes in cup |
| MEN102 Retrieves toy | MEN102 Retrieves toy | MEN102 Retrieves toy |
| MEN104 Uses rod to get toy | MEN104 Uses rod to get toy | MEN104 Uses rod to get toy |
| **X1MTL5 Uses words** | **X1MTL_E Names object** | **X2MTL_E Names object** |
| MEN099 Points to 2 pictures | [MEN099 deleted] | [MEN099 deleted] |
| MEN100 Uses 2 different words | MEN100 Uses 2 different words | MEN100 Uses 2 different words |
| MEN101 Shows shoe | MEN101 Shows shoe | MEN101 Shows shoe |
| MEN106 Uses words to make wants known | MEN106 Uses words to make wants known | MEN106 Uses words to make wants known |
| | MEN110 Names 1 object (new) | MEN110 Names 1 object (new) |
| | **X1MTL_F Receptive vocabulary** | **X2MTL_F Receptive vocabulary** |
| | MEN108 Points to 3 doll parts | MEN108 Points to 3 doll parts |
| | MEN099 Points to 2 pictures | MEN099 Points to 2 pictures |
| | MEN122 Points to 5 pictures | MEN122 Points to 5 pictures |
| | **X1MTL_G Expressive vocabulary** | **X2MTL_G Expressive vocabulary** |
| | MEN111 Combines word/gesture | MEN111 Combines word/gesture |
| | MEN114 Uses s 2-word utterance | MEN114 Uses s 2-word utterance |
| | MEN121 Uses pronouns | MEN121 Uses pronouns |
| | MEN126 Names 3 objects | MEN126 Names 3 objects |
| | MEN133 Names 5 pictures | MEN133 Names 5 pictures |

See note at end of exhibit.

Exhibit 5-1. Changes to 9-month proficiency levels following recalibration using longitudinal 9-month–2-year dataset, original 9-month proficiency level variable names, new 9-month proficiency level variable names, and 2-year proficiency level variable names included within each proficiency: 2001–02 and 2003–04—Continued

| Original 9-month proficiencies in the 9-month data file | 9-month recalibrated proficiencies in the longitudinal 9-month–2-year data file | 2-year proficiencies in the longitudinal 9-month – 2-year data file |
|---|---|---|
|  | **X1MTL_H**<br>**Listening/comprehension**<br>MEN131 Attends to story<br>MEN134 Displays verbal comp.<br>MEN140 Understands 2 prepositions<br>MEN142 Uses multiword utterance in response to book | **X2MTL_H**<br>**Listening/comprehension**<br>MEN131 Attends to story<br>MEN134 Displays verbal comp.<br>MEN140 Understands 2 prepositions<br>MEN142 Uses multiword utterance in response to book |
|  | **X1MTL_I**<br>**Matching/discrimination**<br>MEN125 Matches pictures<br>MEN123 Matches 3 colors<br>MEN137 Matches 4 colors<br>MEN144 Discriminates pictures I<br>MEN151 Discriminates pictures II | **X2MTL_I**<br>**Matching/discrimination**<br>MEN125 Matches pictures<br>MEN123 Matches 3 colors<br>MEN137 Matches 4 colors<br>MEN144 Discriminates pictures I<br>MEN151 Discriminates pictures II |
|  | **X1MTL_J**<br>**Early counting/quantitative**<br>MEN141 Understands concept of 1<br>MEN146 Counts<br>MEN147 Compares masses<br>MEN152 Repeats 3 number sequence<br>MEN159 Counts (stable number order)<br>MEN156 Understands “more”<br>MEN164 Counts (cardinality) | **X2MTL_J**<br>**Early counting/quantitative**<br>MEN141 Understands concept of 1<br>MEN146 Counts<br>MEN147 Compares masses<br>MEN152 Repeats 3 number sequence<br>MEN159 Counts (stable number order)<br>MEN156 Understands “more”<br>MEN164 Counts (cardinality) |
| Motor scale |  |  |
| **X1MTR1 Eye-hand coordination**<br>MOT031 Uses partial thumb opposition to grasp cube<br>MOT049 Uses partial thumb opposition to grasp pellet<br>MOT032 Attempts to secure pellet<br>MOT041 Uses whole hand to grasp pellet | **X1MTR_A Eye-hand coordination**<br>MOT031 Uses partial thumb opposition to grasp cube<br>MOT049 Uses partial thumb opposition to grasp pellet<br>MOT032 Attempts to secure pellet<br>MOT041 Uses whole hand to grasp pellet | **X2MTR_A Eye-hand coordination**<br>MOT031 Uses partial thumb opposition to grasp cube<br>MOT049 Uses partial thumb opposition to grasp pellet<br>MOT032 Attempts to secure pellet<br>MOT041 Uses whole hand to grasp pellet |
| **X1MTR2 Sitting**<br>MOT022 Sits with slight support<br>MOT028 Sits alone momentarily<br>MOT036 Sits alone steadily<br>MOT034 Sits alone for 30 sec.<br>MOT043 Moves forward using prewalking movements<br>MOT051 Moves from sit to creeping position | **X1MTR_B Sitting**<br>MOT022 Sits with slight support<br>MOT028 Sits alone momentarily<br>MOT036 Sits alone steadily<br>MOT034 Sits alone for 30 sec.<br>MOT043 Moves forward using prewalking movements<br>MOT051 Moves from sit to creeping position | **X2MTR_B Sitting**<br>MOT022 Sits with slight support<br>MOT028 Sits alone momentarily<br>MOT036 Sits alone steadily<br>MOT034 Sits alone for 30 sec.<br>MOT043 Moves forward using prewalking movements<br>MOT051 Moves from sit to creeping position |

See note at end of exhibit.

Exhibit 5-1. Changes to 9-month proficiency levels following recalibration using longitudinal 9-month–2-year dataset, original 9-month proficiency level variable names, new 9-month proficiency level variable names, and 2-year proficiency level variable names included within each proficiency: 2001–02 and 2003–04—Continued

| Original 9-month proficiencies in the 9-month data file | 9-month recalibrated proficiencies in the longitudinal 9-month–2-year data file | 2-year proficiencies in the longitudinal 9-month–2-year data file |
|---|---|---|
| **X1MTR3 Pre-walking**<br>MOT044 Supports weight momentarily<br>MOT045 Pulls to standing<br>MOT046 Shifts weight while standing<br>MOT052 Raises self to standing<br>MOT053 Attempts to walk<br>MOT054 Walks sideways while holding onto furniture | **X1MTR_C Pre-walking**<br>MOT044 Supports weight momentarily<br>MOT045 Pulls to standing<br>MOT046 Shifts weight while standing<br>MOT052 Raises self to standing<br>MOT053 Attempts to walk<br>MOT054 Walks sideways while holding onto furniture | **X2MTR_C Pre-walking**<br>MOT044 Supports weight momentarily<br>MOT045 Pulls to standing<br>MOT046 Shifts weight while standing<br>MOT052 Raises self to standing<br>MOT053 Attempts to walk<br>MOT054 Walks sideways while holding onto furniture |
| **X1MTR4 Independent walking**<br>MOT060 Walks with help<br>MOT061 Stands alone<br>MOT059 Stands up I<br>MOT063 Walks alone with good coordination<br>MOT062 Walks alone | **X1MTR_D Stands alone**<br>MOT060 Walks with help<br>MOT061 Stands alone<br>MOT059 Stands up I<br>[MOT063 deleted]<br>MOT062 Walks alone | **X2MTR_D Stands alone**<br>MOT060 Walks with help<br>MOT061 Stands alone<br>MOT059 Stands up I<br>[MOT063 deleted]<br>MOT062 Walks alone |
| **X1MTR5 Balance**<br>(all moved to X2MTR_F)<br>MOT072 Stands on right foot with help<br>MOT065 Squats briefly<br>MOT068 Stands up II<br>MOT073 Stands on left foot with help | **X1MTR_E Skillful walking**<br>MOT063 Walks alone with good coordination<br>MOT067 Walks backward (new)<br>MOT071 Walks sideways (new) | **X2MTR_E Skillful walking**<br>MOT063 Walks alone with good coordination<br>MOT067 Walks backward (new)<br>MOT071 Walks sideways (new) |
|  | **X1MTR_F Balance**<br>MOT065 Squats briefly<br>MOT068 Stands up II<br>MOT072 Stands on right foot with help<br>MOT073 Stands on left foot with help | **X2MTR_F Balance**<br>MOT065 Squats briefly<br>MOT068 Stands up II<br>MOT072 Stands on right foot with help<br>MOT073 Stands on left foot with help |
|  | **X1MTR_G Fine motor control**<br>MOT074 Uses fingerpads to grasp pencil<br>MOT075 Uses hand to hold paper<br>MOT090 Grasps pencil at end nearest point | **X2MTR_G Fine motor control**<br>MOT074 Uses fingerpads to grasp pencil<br>MOT075 Uses hand to hold paper<br>MOT090 Grasps pencil at end nearest point |

See note at end of exhibit.

Exhibit 5-1. Changes to 9-month proficiency levels following recalibration using longitudinal 9-month–2-year dataset, original 9-month proficiency level variable names, new 9-month proficiency level variable names, and 2-year proficiency level variable names included within each proficiency: 2001–02 and 2003–04—Continued

| Original 9-month proficiencies in the 9-month data file | 9-month recalibrated proficiencies in the longitudinal 9-month–2-year data file | 2-year proficiencies in the longitudinal 9-month–2-year data file |
|---|---|---|
| | **X1MTR_H Uses stairs**<br>MOT069 Walks down stairs with help<br>MOT079 Walks up stairs alone, placing both feet on each step<br>MOT080 Walks down stairs alone, placing both feet on each step<br>MOT095 Walks up stairs, alternating feet | **X2MTR_H Uses stairs**<br>MOT069 Walks down stairs with help<br>MOT079 Walks up stairs alone, placing both feet on each step<br>MOT080 Walks down stairs alone, placing both feet on each step<br>MOT095 Walks up stairs, alternating feet |
| | **X1MTR_I Alternating balance**<br>MOT082 Stands alone on right foot<br>MOT083 Stands alone on left foot<br>MOT089 Walks on tiptoe for 4 steps<br>MOT086 Swings leg to kick ball | **X2MTR_I Alternating balance**<br>MOT082 Stands alone on right foot<br>MOT083 Stands alone on left foot<br>MOT089 Walks on tiptoe for 4 steps<br>MOT086 Swings leg to kick ball |
| | **X1MTR_J Motor planning**<br>MOT096 Copies circles<br>MOT093 Manipulates pencil in hand<br>MOT098 Imitates postures<br>MOT088 Laces 3 beads<br>MOT091 Imitates hand movements<br>MOT101 Buttons one button | **X2MTR_J Motor planning**<br>MOT096 Copies circles<br>MOT093 Manipulates pencil in hand<br>MOT098 Imitates postures<br>MOT088 Laces 3 beads<br>MOT091 Imitates hand movements<br>MOT101 Buttons one button |

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Table 5-1. ECLS-B descriptive statistics for BSF-R proficiency level subscales, including proficiency level subscale labels, numbers of items in each subscale, mean scale score, and 9-month and 2-year percentages, in the data file: 2003–04

| Proficiency level subscale | Number of items in subscale | Mean score[1] (publisher) | 9-month percent[2] (ECLS-B) | 2-year percent (ECLS-B) |
|---|---|---|---|---|
| Mental | | | | |
| X1MTL_A: Explores objects | 6 | 53 | 98.9 | 100.0 |
| X1MTL_B: Explores purposefully | 5 | 68 | 87.1 | 100.0 |
| X1MTL_C: Jabbers expressively | 3 | 83 | 41.5 | 99.9 |
| X1MTL_D: Early problem solving | 4 | 98 | 11.1 | 98.5 |
| X1MTL_E: Names object | 4 | 104 | 4.8 | 97.6 |
| X2MTL_F: Receptive vocabulary | 3 | 116 | 1.5 | 84.8 |
| X2MTL_G: Expressive vocabulary | 5 | 126 | 0.3 | 64.5 |
| X2MTL_H: Listening/comprehension | 4 | 139 | 0.1 | 37.3 |
| X2MTL_I: Matching/discrimination | 5 | 140 | 0.2 | 32.6 |
| X2MTL_J: Early counting/quantitative | 7 | 159 | # | 4.2 |
| Motor | | | | |
| X1MTR_A: Demonstrates eye-hand coordination | 4 | 42 | 89.4 | 99.9 |
| X1MTR_B: Sitting | 6 | 42 | 91.4 | 100.0 |
| X1MTR_C: Pre-walking | 6 | 52 | 71.9 | 99.9 |
| X1MTR_D: Stands alone | 4 | 62 | 32.6 | 99.8 |
| X1MTR_E: Skillful walking | 3 | 70 | 18.2 | 92.8 |
| X2MTR_F: Balance | 4 | 74 | 9.2 | 89.7 |
| X2MTR_G: Fine motor control | 3 | 84 | 4.6 | 56.3 |
| X2MTR_H: Uses stairs | 4 | 87 | 3.6 | 48.9 |
| X2MTR_I: Alternating balance | 4 | 90 | 1.5 | 31.0 |
| X2MTR_J: Motor planning | 6 | 99 | 0.5 | 10.8 |

# Rounds to zero.

[1] Scale score, mean value corresponding to 67 percent of total credit possible on each subscale item set. Mean value obtained using BSID-II publisher item calibrations. Scale scores are reported in publisher raw score metric.

[2] Percentage estimate obtained with the weighted ECLS-B sample.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month data collection, 2001–02 and 2-year data collection; 2003–04; publisher standardization data set of the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

For the motor scale, Balance (at 90 percent) and Fine motor control (at 56 percent) are found at a relatively low level of development for 2 years. By contrast, Motor planning (at about 11 percent) is at the upper limits for most 2-year-olds. In the middle range of motor development, Uses stairs (49 percent) and Alternating balance (31 percent) are milestones appropriate for 2-year-olds.

Each of these levels represents a developmental milestone that is a qualitatively different outcome. A qualitative outcome can be scored 1 for mastery and 0 for nonmastery. However, a more informative alternative is available. The IRT subscales reveal how probable it is that a given child can successfully execute each of the tasks belonging to the scale. By averaging scores over tasks, it is possible to calculate the probability of mastering a developmental milestone. Each subscale is treated as a single item in order to estimate the probability of mastery of each skill. The hierarchical nature of skill item sets justifies using the IRT model in this fashion. These items perform much as they would in a Guttman model (Guttman 1944), where a child who is able to complete a given task is expected to have mastered tasks at lower levels of ability; a failure to complete a given task implies nonmastery of items at higher levels of ability. The only difference is that Guttman items are deterministic, whereas IRT items are probabilistic. If the child masters a given milestone, it is highly likely that the child has also mastered all previous milestones.

Probabilities were calculated from IRT true scores after dividing by the total number of items in the subscale. When the resulting probabilities are plotted against the EAP ability estimates on the x axis, they represent the ICC for what is essentially a super-item constructed out of all of the items in the subscale. Users can analyze developmental milestones by examining performance-level probabilities included in the ECLS-B data file.

## 5.6 ECLS-B Data File

The key BSF-R mental and motor scale scores, standard error, and proficiency levels are included in the ECLS-B data file and listed in table 5-2, which provides the variable names, variable labels, and ranges of values for each.

As a final note, the data file includes a variable for BSID age. BSID age is the child’s age at the time of assessment, adjusted for prematurity. BSID-II Mental and Psychomotor Index scores, together with ECLS-B mental and motor age-normed *T* scores, are all based on the child’s age at assessment, corrected for prematurity. This variable was programmed into the computer-assisted personal interviewing (CAPI) portion of the Child Activities section and was generated automatically for all children. The IRT analyses that were conducted on the shorter BSF-R use BSID age to calculate mental and motor *T* scores. This information was obtained from the parent respondent at the time of the child assessments.

Table 5-2. BSF-R mental scale and motor scale variable names, variable labels, and range of values, 2-year data collection: 2003–04

| Variable name | Variable label | Range of values |
|---|---|---|
| X2MTLTSC | ECLS-B mental age-normed *T* scores in N(50, 10) metric | 15.144 – 88.814 |
| X2MTLSCL | Mental scale (IRT true) score in publisher metric | 92.351 – 174.141 |
| X2MTLSSE | Standard error of mental scale (IRT true) score | 2.378 – 8.294 |
| X2MTL_F | Receptive vocabulary | 0.037 – 1.000 |
| X2MTL_G | Expressive vocabulary | 0.005 – 1.000 |
| X2MTL_H | Listening/comprehension | 0.001 – 0.993 |
| X2MTL_I | Matching/discrimination | 0.004 – 0.990 |
| X2MTL_J | Early counting/quantitative | 0.000 – 0.962 |
| X2MTRTSC | ECLS-B motor age-normed *T* scores in N(50, 10) metric | -15.546 – 97.362 |
| X2MTRSCL | Motor scale (IRT true) score in publisher metric | 56.427 – 108.527 |
| X2MTRSSE | Standard error of motor scale (IRT true)score | 1.624 – 5.421 |
| X2MTR_F | Balance | 0.000 – 1.000 |
| X2MTR_G | Fine motor control | 0.013 – 0.991 |
| X2MTR_H | Uses stairs | 0.003 – 0.993 |
| X2MTR_I | Alternating balance | 0.003 – 0.996 |
| X2MTR_J | Motor planning | 0.001 – 0.971 |

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.

As described in the previous chapter, with the availability of BSF-R scores at both 9 months and 2 years, IRT item calibration was conducted with the full set of ECLS-B data, and all observations were rescored using the new item calibrations. The recalibrated 9-month BSF-R scores in the current data file supersede those that were released in the 9-month data file. Analysts who are specifically interested in the 9-month scores should use the new set of recalibrated 9-month scores because these are based on a more consistent scale metric. In addition, as mentioned in section 5.5, the recalibrated 9-month proficiency probability subscales are slightly different from those released in the 9-month data file. For further information about the first release of 9-month BSF-R data, please refer to chapters 2 to 5 of the *ECLS-B, Methodology Report for the Nine-Month Data Collection, Volume 1: Psychometric Characteristics* (NCES 2005-100) (Andreassen and Fletcher 2005).

The variable names and variable labels of the recalibrated 9-month BSF-R scores in this second release are listed in table 5-3.

## 5.7 Average BSF-R Scores and Probabilities by Key Demographic Variables

Table 5-4 and table 5-5 summarize the average scale scores, *T* scores, and probability proficiency levels for the mental scale and motor scale, respectively, for the total sample and for the main grouping variables. These grouping variables are considered to be key factors that are likely to influence children's BSF-R scores. For example, children living at or above the poverty level tend to have higher scores on almost all variables than children living below the poverty level. For this grouping variable, the average 2-year BSF-R mental scale *T* score (X2MTLTSC) was 46.52 for children living below poverty and 50.96 for children living at or above poverty level. The means presented in these tables represent children of all ages, within each grouping variable, in the current data collection.

Table 5-3. Recalibrated 9-month BSF-R mental scale and motor scale variable names, variable labels, and range of values, 2-year data collection: 2003–04

| Variable name | Variable label | Range of values |
|---|---|---|
| X1RMTLS | X1 R MENTAL SCALE SCORE | 32.04 – 131.17 |
| X1RMTLSE | X1 R MENTAL: STAND ERR IRT MENTAL SCALE SCORER | 3.11 – 10.33 |
| X1RMTRS | X1 R MOTOR SCALE SCORE | 21.16 – 87.10 |
| X1RMTRSE | X1 R MOTOR: STAND ERR MOTOR SCALE SCORE | 1.56 – 8.65 |
| X1RMTLT | X1 R MENTAL T-SCORE | 0.09 – 99.16 |
| X1RMTRT | X1 R MOTOR T-SCORE | 2.35 – 83.86 |
| X1MTL_A | X1 MTL PB A: EXPLORES OBJECTS | 0.02 – 1.00 |
| X1MTL_B | X1 MTL PB B: EXPLORES PURPOSEFULLY | 0.00 – 1.00 |
| X1MTL_C | X1 MTL PB C: JABBERS EXPRESSIVELY | 0.00 – 1.00 |
| X1MTL_D | X1 MTL PB D: EARLY PROBLEM SOLVING | 0.00 – 1.00 |
| X1MTL_E | X1 MTL PB E: NAMES OBJECT | 0.00 – 1.00 |
| X1MTR_A | X1 MTR PB A: EYE-HAND COORDINATION | 0.03 – 1.00 |
| X1MTR_B | X1 MTR PB B: SITTING | 0.02 – 1.00 |
| X1MTR_C | X1 MTR PB C: PRE-WALKING | 0.01 – 1.00 |
| X1MTR_D | X1 MTR PB D: STANDS ALONE | 0.00 – 1.00 |
| X1MTR_E | X1 MTR PB E: SKILLFUL WALKING | 0.00 – 0.99 |

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.

Table 5-4. Weighted means (and standard deviations) of the BSF-R mental scale and mental probability scores by key demographic variables, 2-year data collection: 2003–04

| | | BSF-R mental scale mean scores | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Mental *T* score[1] (X2MTLTSC) | Mental scale score (X2MTLSCL) | Receptive vocabulary (X2MTL_F) | Expressive vocabulary (X2MTL_G) | Listening/ comprehension (X2MTL_H) | Matching/ discrimination (X2MTL_I) | Early counting/ prequantitative (X2MTL_J) |
| Total sample | 8,900 | 50.00 | 127.09 | 0.85 | 0.65 | 0.37 | 0.33 | 0.04 |
| | | (10.00) | (10.65) | (0.19) | (0.27) | (0.22) | (0.21) | (0.09) |
| Child's race/ethnicity[2] | | | | | | | | |
| White | 3,850 | 52.51 | 129.64 | 0.89 | 0.71 | 0.43 | 0.38 | 0.06 |
| | | (9.77) | (10.32) | (0.16) | (0.25) | (0.22) | (0.22) | (0.10) |
| Black | 1,400 | 47.28 | 124.10 | 0.80 | 0.57 | 0.31 | 0.27 | 0.03 |
| | | (9.46) | (10.31) | (0.22) | (0.27) | (0.21) | (0.19) | (0.06) |
| Hispanic, race specified | 1,250 | 46.84 | 123.92 | 0.80 | 0.56 | 0.30 | 0.26 | 0.02 |
| | | (9.31) | (10.10) | (0.21) | (0.27) | (0.20) | (0.19) | (0.06) |
| Hispanic, no race specified | 550 | 45.43 | 122.74 | 0.78 | 0.53 | 0.28 | 0.24 | 0.02 |
| | | (8.93) | (9.77) | (0.21) | (0.26) | (0.20) | (0.18) | (0.06) |
| Asian | 900 | 49.23 | 126.51 | 0.83 | 0.63 | 0.36 | 0.32 | 0.04 |
| | | (10.63) | (11.31) | (0.22) | (0.28) | (0.23) | (0.21) | (0.10) |
| Native Hawaiian, Pacific Islander | 50 | 46.50 | 122.92 | 0.79 | 0.55 | 0.28 | 0.24 | 0.01 |
| | | (8.25) | (9.07) | (0.21) | (0.25) | (0.17) | (0.15) | (0.05) |
| American Indian, Alaska Native | 250 | 45.10 | 122.10 | 0.76 | 0.52 | 0.27 | 0.23 | 0.02 |
| | | (9.31) | (10.62) | (0.24) | (0.28) | (0.20) | (0.18) | (0.06) |
| More than 1 race | 700 | 50.00 | 126.94 | 0.85 | 0.65 | 0.37 | 0.32 | 0.04 |
| | | (9.48) | (10.18) | (0.19) | (0.27) | (0.22) | (0.20) | (0.07) |
| Poverty status | | | | | | | | |
| Below poverty threshold | 1,950 | 46.52 | 123.67 | 0.80 | 0.56 | 0.30 | 0.26 | 0.02 |
| | | (9.15) | (9.99) | (0.21) | (0.27) | (0.20) | (0.18) | (0.06) |
| At or above poverty threshold | 6,950 | 50.96 | 128.03 | 0.86 | 0.67 | 0.39 | 0.34 | 0.05 |
| | | (10.01) | (10.63) | (0.18) | (0.27) | (0.22) | (0.21) | (0.09) |
| Child's sex | | | | | | | | |
| Male | 4,550 | 48.30 | 125.33 | 0.82 | 0.60 | 0.34 | 0.29 | 0.03 |
| | | (9.99) | (10.70) | (0.21) | (0.28) | (0.22) | (0.20) | (0.08) |
| Female | 4,350 | 51.78 | 128.93 | 0.88 | 0.69 | 0.41 | 0.36 | 0.05 |
| | | (9.70) | (10.27) | (0.16) | (0.25) | (0.22) | (0.21) | (0.10) |

See notes at end of table.

Table 5-4. Weighted means (and standard deviations) of the BSF-R mental scale and mental probability scores by key demographic variables, 2-year data collection: 2003–04—Continued

| | | BSF-R mental scale mean scores | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Mental *T* score[1] (X2MTLTSC) | Mental scale score (X2MTLSCL) | Receptive vocabulary (X2MTL_F) | Expressive vocabulary (X2MTL_G) | Listening/ comprehension (X2MTL_H) | Matching/ discrimination (X2MTL_I) | Early counting/ prequantitative (X2MTL_J) |
| Birth weight | | | | | | | | |
| Normal | 6,550 | 50.20 | 127.49 | 0.86 | 0.66 | 0.38 | 0.33 | 0.04 |
| | | (9.97) | (10.52) | (0.19) | (0.27) | (0.22) | (0.21) | (0.09) |
| Moderately low | 1,400 | 47.83 | 123.16 | 0.78 | 0.55 | 0.29 | 0.25 | 0.02 |
| | | (9.99) | (10.78) | (0.23) | (0.28) | (0.21) | (0.19) | (0.07) |
| Very low | 950 | 46.06 | 117.06 | 0.64 | 0.38 | 0.18 | 0.16 | 0.01 |
| | | (9.68) | (10.43) | (0.27) | (0.27) | (0.17) | (0.15) | (0.03) |
| Child’s age at assessment | | | | | | | | |
| 21 months and under | # | 58.53 | 125.45 | 0.89 | 0.63 | 0.33 | 0.26 | 0.01 |
| | | (3.71) | (3.08) | (0.07) | (0.11) | (0.08) | (0.06) | (0.00) |
| 22–23 months | 850 | 50.34 | 123.93 | 0.79 | 0.57 | 0.31 | 0.27 | 0.03 |
| | | (10.20) | (10.79) | (0.23) | (0.28) | (0.21) | (0.19) | (0.06) |
| 24–25 months | 6,850 | 50.37 | 126.94 | 0.85 | 0.64 | 0.37 | 0.32 | 0.04 |
| | | (9.79) | (10.40) | (0.19) | (0.27) | (0.22) | (0.21) | (0.08) |
| 26–27 months | 950 | 47.64 | 129.83 | 0.89 | 0.71 | 0.43 | 0.38 | 0.06 |
| | | (10.20) | (10.49) | (0.16) | (0.25) | (0.22) | (0.22) | (0.11) |
| 28 months and over | 250 | 46.26 | 134.56 | 0.93 | 0.79 | 0.52 | 0.47 | 0.12 |
| | | (12.25) | (12.13) | (0.11) | (0.22) | (0.24) | (0.25) | (0.19) |
| Mother’s age (in years) | | | | | | | | |
| 19 and under | 300 | 47.24 | 124.34 | 0.81 | 0.57 | 0.31 | 0.26 | 0.02 |
| | | (8.77) | (9.30) | (0.18) | (0.26) | (0.20) | (0.18) | (0.06) |
| 20–29 | 3,950 | 48.78 | 125.93 | 0.83 | 0.62 | 0.35 | 0.30 | 0.03 |
| | | (9.62) | (10.32) | (0.19) | (0.27) | (0.22) | (0.20) | (0.08) |
| 30–39 | 3,900 | 51.51 | 128.59 | 0.87 | 0.68 | 0.41 | 0.36 | 0.05 |
| | | (10.23) | (10.85) | (0.19) | (0.27) | (0.23) | (0.22) | (0.10) |
| 40 and over | 700 | 50.51 | 127.32 | 0.85 | 0.66 | 0.38 | 0.33 | 0.04 |
| | | (10.23) | (10.91) | (0.21) | (0.27) | (0.22) | (0.21) | (0.09) |

See notes at end of table.

Table 5-4. Weighted means (and standard deviations) of the BSF-R mental scale and mental probability scores by key demographic variables, 2-year data collection: 2003–04—Continued

| | | BSF-R mental scale mean scores | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Mental *T* score[1] (X2MTLTSC) | Mental scale score (X2MTLSCL) | Receptive vocabulary (X2MTL_F) | Expressive vocabulary (X2MTL_G) | Listening/ comprehension (X2MTL_H) | Matching/ discrimination (X2MTL_I) | Early counting/ prequantitative (X2MTL_J) |
| Mother's race/ethnicity[2] | | | | | | | | |
| White | 4,200 | 52.36 | 129.48 | 0.89 | 0.71 | 0.42 | 0.37 | 0.05 |
| | | (9.69) | (10.26) | (0.16) | (0.25) | (0.22) | (0.21) | (0.10) |
| Black | 1,450 | 47.44 | 124.24 | 0.80 | 0.58 | 0.31 | 0.27 | 0.03 |
| | | (9.47) | (10.33) | (0.22) | (0.27) | (0.21) | (0.19) | (0.06) |
| Hispanic, race specified | 1,500 | 45.93 | 123.09 | 0.78 | 0.54 | 0.29 | 0.24 | 0.02 |
| | | (9.14) | (9.94) | (0.21) | (0.27) | (0.20) | (0.18) | (0.06) |
| Hispanic, no race specified | 50 | 46.83 | 124.50 | 0.78 | 0.58 | 0.33 | 0.29 | 0.04 |
| | | (10.61) | (12.63) | (0.28) | (0.32) | (0.24) | (0.22) | (0.09) |
| Asian | 1,050 | 49.34 | 126.56 | 0.84 | 0.63 | 0.37 | 0.32 | 0.04 |
| | | (10.49) | (11.19) | (0.21) | (0.28) | (0.22) | (0.21) | (0.09) |
| Native Hawaiian, Pacific Islander | 50 | 48.28 | 125.17 | 0.81 | 0.58 | 0.33 | 0.29 | 0.05 |
| | | (11.03) | (11.78) | (0.22) | (0.28) | (0.23) | (0.23) | (0.10) |
| American Indian, Alaska Native | 300 | 45.26 | 122.37 | 0.77 | 0.52 | 0.28 | 0.24 | 0.02 |
| | | (9.26) | (10.21) | (0.23) | (0.28) | (0.20) | (0.18) | (0.05) |
| More than 1 race | 250 | 49.32 | 126.56 | 0.84 | 0.62 | 0.36 | 0.31 | 0.04 |
| | | (10.34) | (10.74) | (0.19) | (0.27) | (0.23) | (0.22) | (0.09) |

See notes at end of table.

Table 5-4. Weighted means (and standard deviations) of the BSF-R mental scale and mental probability scores by key demographic variables, 2-year data collection: 2003–04—Continued

| | | BSF-R mental scale mean scores | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Mental *T* score[1] (X2MTLTSC) | Mental scale score (X2MTLSCL) | Receptive vocabulary (X2MTL_F) | Expressive vocabulary (X2MTL_G) | Listening/ comprehension (X2MTL_H) | Matching/ discrimination (X2MTL_I) | Early counting/ prequantitative (X2MTL_J) |
| Mother's education | | | | | | | | |
| 8th grade or below | 400 | 44.23 | 121.74 | 0.77 | 0.51 | 0.25 | 0.22 | 0.01 |
| | | (7.90) | (9.09) | (0.21) | (0.25) | (0.18) | (0.16) | (0.04) |
| 9–12th grades | 1,800 | 47.21 | 124.36 | 0.81 | 0.58 | 0.31 | 0.27 | 0.03 |
| | | (9.10) | (9.72) | (0.20) | (0.26) | (0.20) | (0.18) | (0.07) |
| High school diploma | 1,900 | 48.78 | 125.71 | 0.83 | 0.61 | 0.35 | 0.30 | 0.03 |
| | | (9.78) | (10.59) | (0.21) | (0.27) | (0.22) | (0.20) | (0.08) |
| Vocational/technical | 200 | 50.39 | 127.45 | 0.86 | 0.65 | 0.38 | 0.33 | 0.04 |
| | | (9.54) | (10.16) | (0.17) | (0.25) | (0.22) | (0.21) | (0.09) |
| Some college | 2,150 | 50.47 | 127.56 | 0.86 | 0.66 | 0.38 | 0.34 | 0.04 |
| | | (9.73) | (10.36) | (0.19) | (0.27) | (0.22) | (0.21) | (0.08) |
| Bachelor's degree | 1,450 | 53.33 | 130.38 | 0.90 | 0.73 | 0.44 | 0.39 | 0.06 |
| | | (9.79) | (10.43) | (0.16) | (0.25) | (0.22) | (0.22) | (0.11) |
| Graduate school (no degree) | 150 | 55.12 | 132.33 | 0.92 | 0.78 | 0.49 | 0.44 | 0.08 |
| | | (9.58) | (10.23) | (0.16) | (0.23) | (0.21) | (0.21) | (0.11) |
| Master's degree | 600 | 56.18 | 133.30 | 0.92 | 0.79 | 0.51 | 0.46 | 0.09 |
| | | (10.19) | (10.64) | (0.14) | (0.23) | (0.22) | (0.23) | (0.13) |
| Doctoral/professional degree | 200 | 55.38 | 132.75 | 0.92 | 0.78 | 0.50 | 0.45 | 0.08 |
| | | (9.66) | (10.30) | (0.15) | (0.22) | (0.21) | (0.22) | (0.12) |

# Rounds to zero

[1] Mental *T* scores are age-adjusted.

[2] Race categories exclude Hispanic origin unless specified.

NOTE: Results were obtained by using the sampling child weight W2C0, however cell counts are unweighted to demonstrate better the distribution in the ECLS-B. Standard deviations appear in parentheses. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

Table 5-5. Weighted means (and standard deviations) of the BSF-R motor scale and motor probability scores by key demographic variables, 2-year data collection: 2003–04

| | | BSF-R motor scale mean scores | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Motor *T* score[1] (X2MTRTSC) | Motor scale score (X2MTRSCL) | Balance (X2MTR_F) | Fine motor control (X2MTR_G) | Uses stairs (X2MTR_H) | Alternating balance (X2MTR_I) | Motor planning (X2MTR_J) |
| Total sample | 8,850 | 50.00 | 81.47 | 0.90 | 0.56 | 0.49 | 0.31 | 0.11 |
| | | (10.00) | (5.07) | (0.15) | (0.18) | (0.16) | (0.17) | (0.08) |
| Child's race/ethnicity[2] | | | | | | | | |
| White | 3,800 | 50.20 | 81.53 | 0.90 | 0.56 | 0.49 | 0.31 | 0.11 |
| | | (9.93) | (4.95) | (0.14) | (0.18) | (0.15) | (0.17) | (0.08) |
| Black | 1,400 | 51.65 | 82.21 | 0.91 | 0.59 | 0.51 | 0.34 | 0.12 |
| | | (9.87) | (5.10) | (0.14) | (0.18) | (0.16) | (0.18) | (0.09) |
| Hispanic, race specified | 1,250 | 48.57 | 80.84 | 0.88 | 0.54 | 0.47 | 0.29 | 0.10 |
| | | (9.96) | (5.15) | (0.17) | (0.19) | (0.16) | (0.17) | (0.08) |
| Hispanic, no race specified | 550 | 49.09 | 81.27 | 0.89 | 0.55 | 0.48 | 0.31 | 0.11 |
| | | (10.39) | (5.43) | (0.16) | (0.20) | (0.17) | (0.19) | (0.09) |
| Asian | 900 | 49.53 | 81.35 | 0.89 | 0.56 | 0.48 | 0.31 | 0.11 |
| | | (10.09) | (5.18) | (0.15) | (0.19) | (0.16) | (0.18) | (0.09) |
| Native Hawaiian, Pacific Islander | 50 | 50.06 | 81.15 | 0.90 | 0.55 | 0.48 | 0.29 | 0.10 |
| | | (7.53) | (4.44) | (0.12) | (0.16) | (0.14) | (0.16) | (0.08) |
| American Indian, Alaska Native | 250 | 49.61 | 81.38 | 0.89 | 0.56 | 0.49 | 0.31 | 0.11 |
| | | (9.91) | (5.19) | (0.14) | (0.19) | (0.16) | (0.18) | (0.08) |
| More than 1 race | 650 | 49.76 | 81.24 | 0.89 | 0.55 | 0.48 | 0.30 | 0.10 |
| | | (9.71) | (4.99) | (0.14) | (0.19) | (0.16) | (0.17) | (0.08) |
| Poverty status | | | | | | | | |
| Below poverty threshold | 1,950 | 49.25 | 81.24 | 0.89 | 0.56 | 0.48 | 0.30 | 0.11 |
| | | (10.09) | (5.17) | (0.16) | (0.19) | (0.16) | (0.17) | (0.08) |
| At or above poverty threshold | 6,900 | 50.21 | 81.53 | 0.90 | 0.57 | 0.49 | 0.31 | 0.11 |
| | | (9.96) | (5.04) | (0.14) | (0.18) | (0.16) | (0.17) | (0.08) |
| Child's sex | | | | | | | | |
| Male | 4,500 | 49.28 | 81.11 | 0.89 | 0.55 | 0.48 | 0.30 | 0.10 |
| | | (9.99) | (5.09) | (0.16) | (0.19) | (0.16) | (0.17) | (0.07) |
| Female | 4,350 | 50.76 | 81.85 | 0.91 | 0.58 | 0.50 | 0.32 | 0.11 |
| | | (9.95) | (5.02) | (0.13) | (0.18) | (0.16) | (0.18) | (0.09) |

See notes at end of table.

Table 5-5. Weighted means (and standard deviations) of the BSF-R motor scale and motor probability scores by key demographic variables, 2-year data collection: 2003–04—Continued

| | | BSF-R motor scale mean scores | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Motor<br>*T* score[1]<br>(X2MTRTSC) | Motor<br>scale score<br>(X2MTRSCL) | Balance<br>(X2MTR_F) | Fine motor<br>control<br>(X2MTR_G) | Uses stairs<br>(X2MTR_H) | Alternating<br>balance<br>(X2MTR_I) | Motor<br>planning<br>(X2MTR_J) |
| Birth weight | | | | | | | | |
| Normal | 6,500 | 50.12 | 81.65 | 0.90 | 0.57 | 0.49 | 0.32 | 0.11 |
| | | (9.93) | (4.97) | (0.14) | (0.18) | (0.15) | (0.17) | (0.08) |
| Moderately low | 1,400 | 48.72 | 79.77 | 0.85 | 0.50 | 0.44 | 0.26 | 0.09 |
| | | (10.53) | (5.52) | (0.19) | (0.20) | (0.16) | (0.16) | (0.07) |
| Very low | 900 | 47.40 | 76.73 | 0.75 | 0.39 | 0.35 | 0.18 | 0.06 |
| | | (11.65) | (5.81) | (0.25) | (0.20) | (0.16) | (0.14) | (0.05) |
| Child's age at assessment | | | | | | | | |
| 21 months and under | # | 62.91 | 81.73 | 0.93 | 0.59 | 0.50 | 0.30 | 0.10 |
| | | (6.68) | (3.06) | (0.11) | (0.13) | (0.09) | (0.08) | (0.03) |
| 22–23 months | 850 | 52.75 | 80.59 | 0.87 | 0.53 | 0.46 | 0.28 | 0.10 |
| | | (10.50) | (5.29) | (0.17) | (0.20) | (0.16) | (0.17) | (0.08) |
| 24–25 months | 6,800 | 50.21 | 81.21 | 0.90 | 0.56 | 0.48 | 0.30 | 0.10 |
| | | (9.55) | (4.82) | (0.15) | (0.18) | (0.15) | (0.16) | (0.07) |
| 26–27 months | 950 | 46.46 | 83.12 | 0.92 | 0.62 | 0.54 | 0.38 | 0.14 |
| | | (10.59) | (5.31) | (0.13) | (0.18) | (0.16) | (0.20) | (0.10) |
| 28 months and over | 250 | 45.20 | 86.54 | 0.96 | 0.71 | 0.64 | 0.51 | 0.23 |
| | | (12.85) | (6.37) | (0.11) | (0.18) | (0.18) | (0.24) | (0.17) |
| Mother's age (in years) | | | | | | | | |
| 19 and under | 300 | 49.11 | 81.11 | 0.88 | 0.55 | 0.48 | 0.30 | 0.11 |
| | | (10.73) | (5.56) | (0.17) | (0.20) | (0.17) | (0.19) | (0.09) |
| 20–29 | 3,950 | 49.74 | 81.42 | 0.90 | 0.56 | 0.49 | 0.31 | 0.11 |
| | | (9.85) | (5.07) | (0.15) | (0.18) | (0.16) | (0.17) | (0.08) |
| 30–39 | 3,900 | 50.47 | 81.64 | 0.90 | 0.57 | 0.49 | 0.32 | 0.11 |
| | | (10.06) | (5.01) | (0.14) | (0.18) | (0.16) | (0.17) | (0.08) |
| 40 and over | 700 | 49.23 | 80.89 | 0.88 | 0.54 | 0.47 | 0.29 | 0.10 |
| | | (10.17) | (5.18) | (0.16) | (0.19) | (0.16) | (0.17) | (0.08) |

See notes at end of table.

Table 5-5. Weighted means (and standard deviations) of the BSF-R motor scale and motor probability scores by key demographic variables, 2-year data collection: 2003–04—Continued

| | | BSF-R motor scale mean scores | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Motor *T* score[1] (X2MTRTSC) | Motor scale score (X2MTRSCL) | Balance (X2MTR_F) | Fine motor control (X2MTR_G) | Uses stairs (X2MTR_H) | Alternating balance (X2MTR_I) | Motor planning (X2MTR_J) |
| Mother's race/ethnicity[2] | | | | | | | | |
| White | 4,200 | 50.22 | 81.53 | 0.90 | 0.56 | 0.49 | 0.31 | 0.11 |
| | | (9.86) | (4.95) | (0.14) | (0.18) | (0.15) | (0.17) | (0.08) |
| Black | 1,400 | 51.81 | 82.27 | 0.91 | 0.59 | 0.51 | 0.34 | 0.12 |
| | | (9.88) | (5.11) | (0.14) | (0.18) | (0.16) | (0.18) | (0.09) |
| Hispanic, race specified | 1,500 | 48.62 | 80.94 | 0.88 | 0.54 | 0.47 | 0.29 | 0.10 |
| | | (10.20) | (5.29) | (0.17) | (0.19) | (0.16) | (0.17) | (0.08) |
| Hispanic, no race specified | 50 | 48.73 | 81.32 | 0.90 | 0.55 | 0.48 | 0.31 | 0.11 |
| | | (8.81) | (5.00) | (0.13) | (0.19) | (0.16) | (0.18) | (0.07) |
| Asian | 1,050 | 49.27 | 81.17 | 0.89 | 0.55 | 0.48 | 0.30 | 0.10 |
| | | (10.22) | (5.15) | (0.15) | (0.19) | (0.16) | (0.18) | (0.08) |
| Native Hawaiian, Pacific Islander | 50 | 51.51 | 82.11 | 0.91 | 0.59 | 0.51 | 0.33 | 0.12 |
| | | (8.98) | (4.94) | (0.13) | (0.18) | (0.15) | (0.17) | (0.09) |
| American Indian, Alaska Native | 300 | 49.07 | 81.17 | 0.89 | 0.55 | 0.48 | 0.30 | 0.10 |
| | | (9.48) | (4.86) | (0.13) | (0.18) | (0.15) | (0.17) | (0.07) |
| More than 1 race | 250 | 47.22 | 80.19 | 0.87 | 0.52 | 0.45 | 0.27 | 0.09 |
| | | (10.25) | (4.92) | (0.16) | (0.19) | (0.15) | (0.16) | (0.06) |

See notes at end of table.

Table 5-5. Weighted means (and standard deviations) of the BSF-R motor scale and motor probability scores by key demographic variables, 2-year data collection: 2003–04—Continued

| Characteristic | Number | BSF-R motor scale mean scores | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Motor *T* score[1] (X2MTRTSC) | Motor scale score (X2MTRSCL) | Balance (X2MTR_F) | Fine motor control (X2MTR_G) | Uses stairs (X2MTR_H) | Alternating balance (X2MTR_I) | Motor planning (X2MTR_J) |
| Mother's education | | | | | | | | |
| 8th grade or below | 400 | 47.65 | 80.73 | 0.88 | 0.54 | 0.47 | 0.28 | 0.10 |
| | | (9.41) | (5.05) | (0.16) | (0.19) | (0.16) | (0.17) | (0.07) |
| 9–12th grades | 1,800 | 49.47 | 81.32 | 0.89 | 0.56 | 0.48 | 0.31 | 0.11 |
| | | (9.95) | (5.11) | (0.16) | (0.19) | (0.16) | (0.17) | (0.08) |
| High school diploma | 1,900 | 49.79 | 81.31 | 0.89 | 0.56 | 0.48 | 0.31 | 0.11 |
| | | (9.95) | (5.07) | (0.15) | (0.19) | (0.16) | (0.17) | (0.08) |
| Vocational/technical | 150 | 51.91 | 82.39 | 0.92 | 0.59 | 0.52 | 0.34 | 0.12 |
| | | (9.98) | (5.05) | (0.12) | (0.18) | (0.16) | (0.19) | (0.09) |
| Some college | 2,150 | 49.87 | 81.39 | 0.90 | 0.56 | 0.49 | 0.31 | 0.11 |
| | | (10.17) | (5.04) | (0.14) | (0.18) | (0.16) | (0.17) | (0.08) |
| Bachelor's degree | 1,450 | 51.01 | 81.84 | 0.91 | 0.57 | 0.50 | 0.32 | 0.11 |
| | | (10.05) | (5.10) | (0.14) | (0.18) | (0.16) | (0.18) | (0.09) |
| Graduate school (no degree) | 150 | 49.80 | 81.30 | 0.90 | 0.56 | 0.48 | 0.30 | 0.10 |
| | | (8.59) | (4.39) | (0.13) | (0.17) | (0.14) | (0.15) | (0.06) |
| Master's degree | 600 | 51.47 | 82.06 | 0.91 | 0.58 | 0.51 | 0.33 | 0.12 |
| | | (9.96) | (5.07) | (0.13) | (0.18) | (0.16) | (0.18) | (0.09) |
| Doctoral/professional degree | 200 | 51.47 | 82.25 | 0.91 | 0.59 | 0.51 | 0.34 | 0.12 |
| | | (9.56) | (4.91) | (0.13) | (0.18) | (0.15) | (0.18) | (0.08) |

# Rounds to zero.

[1] Motor *T* scores are age-adjusted.

[2] Race categories exclude Hispanic origin unless specified.

NOTE: Results were obtained by applying the sampling child weight W2C0, however cell counts are unweighted to demonstrate better the distribution in the ECLS-B. Standard deviations appear in parentheses. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

# 6. TWO BAGS TASK IN THE ECLS-B

This section discusses the rationale for the transition from the Nursing Child Assessment Teaching Scale (NCATS) at 9 months to the Two Bags Task at 2 years. This is followed by a description of the in-home administration of the Two Bags Task, as well as the quality control procedures that were undertaken to ensure that the data obtained were of the highest quality, including training the interviewers, training the trainers on the coding system, and training the coders. In addition, a summary of coder reliability and Cronbach's alpha for the Two Bags subscales is presented, followed by a comparison of the 2-year Two Bags Task rating scales with the 9-month NCATS scale scores and descriptives for the Two Bags Task by key demographic grouping variables.

## 6.1 Technical Review Panel Advice

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.

Consideration of the above principles led to the decision, strongly endorsed by Technical Review Panel (TRP) advisors, to include a direct measure of parent-child interaction. Parent-child interaction is a key aspect of children's early experiences known to predict subsequent child outcomes. To capture the full breadth of young children's functioning, it was important to include a direct measure of their socioemotional functioning. During infancy and toddlerhood, socioemotional functioning is easiest to assess during mother-child interaction, because this interaction provides a context within which the child's emotional functioning can be elicited. This approach is supported by multiple lines of research in developmental psychology, including attunement (Stern 1985), intersubjectivity (Trevarthen and Aitken 2001), social referencing (Walker-Andrews 1998), and emotion regulation (Miller et al. 2002).

TRP advisors advocated the inclusion of a direct observational measure of the parent's and child's behaviors in order to obtain information about the quality of children's interactions that may influence their readiness for school. TRP members further advised that the NCATS of the Nursing Child Assessment Satellite Training (NCAST) would be useful for the early rounds of the ECLS-B for several reasons. It was used in the Early Head Start Research and Evaluation project, a study with a large sample of very young children, enabling a comparison of the results from the two studies; it has a standardized training that would ensure consistency of coding; it could be used at both the 9- and 18-month data collections, which would further the goal of obtaining repeated measures to examine growth; and it can be coded from videotape, thereby reducing burden to interviewers who would simply videotape the interaction and not code during administration.

A different observational measure of parent-child interaction needed to be implemented at the 30-month and later data collections because the developer of the NCATS did not endorse its use beyond 24 months of age. For the purposes of the ECLS-B 2-year data collection, the TRP members suggested the Three Bags Task as a viable measure of parent-child interaction, because it is one of the few coding systems that can be used in large-scale studies, has excellent training materials, good psychometric properties, and, while brief, produces robust scores predictive of later growth in both cognitive and socioemotional domains. As design of the 30-month data collection progressed, it was decided to implement the Three Bags Task, which has been used with success in other large-scale studies, including the Early Head Start Research and Evaluation Project (with a national sample of approximately 3,000 very young children) sponsored by the Administration on Children, Youth, and Families (ACYF), and the National Institute of Child Health and Human Development (NICHD) Early Child Care Study, a project involving a consortium of academic and social policy researchers (NICHD Early Child Care Research Network 2004, 2005).

The Three Bags Task is a semi-structured activity completed by the parent and child in interaction. Because the Three Bags Task requires at least 15 minutes for the dyad to complete, the activity was shortened to just two bags/activities, which could be completed in 10 minutes, with the provision that one activity should be a joint book reading activity. During this 10-minute task, the parent-child dyad is asked to play with two different sets of toys, each placed within a separate numbered bag. In the 2-year ECLS-B, bag number 1 contained a set of dishes, and bag number 2 contained a children’s picture book, *Good Night, Gorilla,* by P. Rathmann (1994). The dyad was told that they had 10 minutes to play with the two bags, the only restriction being that they had to play with the bags in numerical order. The parent and child were videotaped while they engaged in the activities. The videotapes were sent back to Westat, where staff trained on the rating scale rated the parent on six global scales and the child on three global scales.

The Two Bags Task rating scales include six parent rating scales and three child rating scales. The scales are on a 7-point Likert-type rating scale that ranged from very low (1) to very high (7). Each rating level is well described in the coding manual with specific examples to illustrate the concept and target behaviors.

To code a videotape, the coder watched the videotape and observed the target behaviors, making notes that would help rate the items. When the videotape was finished, the coder rated each item on the basis of observations made while watching the videotape.

The six parent rating scales include the following:

- Parental Sensitivity (variable name C2SENSTV): This scale focuses on how the parent observes and responds to the child's cues (including gestures, expressions, and signals), both when the child is distressed and not distressed. The key defining characteristic of parental sensitivity is that the parent's response is child-centered. Sensitive parenting involves "tuning-in" to the child and manifesting awareness of the child's needs, moods, interests, and capabilities.

- Parental Intrusiveness (variable name C2NTRUSV): This scale reflects the degree to which the parent controls the child rather than recognizes and respects the validity of the child's perspective. Intrusive interactions are adult-centered rather than child-centered and involve imposing the parent's agenda on the child despite the child's protest or defensiveness. Extreme intrusiveness can be seen as over-control to the point where the child's autonomy is minimized or rejected. The key characteristic is that the intrusiveness is seen from the point of view of the child and careful observation of the child's reaction to the intrusiveness is required.

- Parental Stimulation of Cognitive Development (variable name C2COGDEV): This scale focuses on the parent's effortful teaching to enhance perceptual, cognitive, and language development. A stimulating parent is aware of the child's developmental level and aims to bring the child to the next level. If the topic or method of stimulation is not matched to or slightly above the child's developmental level or interest, then the parent's behavior is not seen as stimulating cognitive development.

- Parental Positive Regard (variable name C2POSRGD): This scale assesses the parent's expression of love, respect, and admiration for the child. Positive regard is seen in the way the parent listens, watches attentively, and looks into the child's face when talking to him/her. Parents who give praise without a warm tone as well as those who do not praise when the opportunity presents itself would not receive the highest score.

- Parental Negative Regard (variable name C2NEGRGD): This scale reflects the parent's expression of discontent with, anger toward, disapproval of, or rejection of the child. The key is to score parental negative regard from the point of view of the child, and it should be scored independently of the parent's positive behaviors captured in the positive regard scale.

- Parental Detachment (variable name C2DETACH): This scale measures the parent's awareness of, attention to, and engagement with the child. This includes both the extent to which the parent interacts with the child (i.e., the amount of interaction) and the way in which the parent interacts with the child (i.e., the quality of interaction). Detachment can take the form of being consistently inattentive, being inconsistently attentive, or interacting with the child in a perfunctory or indifferent manner.

The three scales that assess children's behaviors include the following:

- Child Engagement of Parent (variable name C2ENGPRT): This scale reflects the extent to which the child shows, initiates, and maintains interaction with the parent, and the extent to which the child communicates positive regard or positive affect to the parent. At the higher end of the scale, the child expresses sustained positive affects toward the parent (through smiling, laughter, etc.) and frequently looks at and attempts to interact with the parent. At the lower end of the scale, the child displays no affect with the parent or ignores or overtly rejects the parent.

- Child Sustained Attention (variable name C2STNATT): This scale assesses the child's ability to sustain attention to and involvement with objects. A child low on sustained attention could seem apathetic, bored, distracted, distressed, or aimless while a child high on sustained attention is able to focus attention when playing with an object and appears involved in what he/she is doing.

- Child Negativity Toward Parent (variable name C2NEGPRT): This scale measures the degree to which the child shows anger, hostility, or dislike toward the parent. At the high end, the child is repeatedly and overtly angry with the parent. The important point is that at this age, the child may express negativity toward the parent by hitting an object, the floor, or him/herself by pushing the parent away, by throwing a toy, or by using a negative expression to communicate that he/she wants or does not want something ("No!"). Therefore, the context of the negative expression should be taken into account when determining the extent to which it is directed toward the parent.

## 6.2 Rationale for Transition to Two Bags Task

The combining of the 18- and 30-month data collections into a single 2-year data collection necessitated a decision about whether to use the NCATS or the Two Bags Task. Had the 18-month data collection gone forward as planned, there would be no question that the NCATS would be included at 18 months and the Two Bags Task at 30 months and at preschool. However, the switch to 2 years shifted the considerations somewhat with regard to obtaining continuity of measurement. For one, a decision had to be made about which measure, NCATS or Two Bags, would provide the repeated measure for estimating growth over time: NCATS at 9 months and 2 years, and Two Bags at preschool, or NCATS at 9 months only and Two Bags at 2 years and preschool. The decision was made to administer the NCATS at 9 months and the Two Bags at 2 years and preschool.

Several factors contributed to the decision about which combination of assessments to use. First, 2 years is at the upper limit of the age range for which empirical data support the use of the NCATS, according to its developer, Dr. Kathryn Barnard of the University of Washington. She did not encourage using the NCATS beyond that age because most research using the NCATS has been concentrated on children up to about 2 years of age, with relatively fewer studies on children older than 2 years of age. Consequently, it is not clear that the NCATS can reliably measure parent-child interactions for children older than age 2 (i.e., 24 months). The lack of a research base for the NCATS beyond 2 years would not be a problem if all the children in the ECLS-B were seen promptly within the predetermined "ideal window" of 2 years +/- 4 weeks. However, experience during the 9-month collection when some of the children were seen many months later demonstrated that it was unrealistic to expect that all home visits would be completed within this window. Because only one observational measure could reasonably be used at 2 years, the ECLS-B had to select a different measure if there were no empirical support for the NCATS norms beyond 2 years.

Second, the Two Bags task has the advantage that it can use a parent-child joint book reading activity as one of the tasks. This would give the ECLS-B the opportunity to obtain a direct observational measure of mother's and child's language use and literacy behaviors, an important consideration for a study examining the aspects of children's early experiences that prepare them for later school entry and sustained school achievement. Indeed, Hart and Risley (1995), have built a strong argument for the effects of early experiences on children's later outcomes, in particular, that the amount of time parents spend talking to their children in the early years of life directly influences children's future school achievement.

The third consideration in choosing an assessment was ease of administration and coding, as well as cost. It is quite expensive to obtain videotaped interactions of parent and child. Several coding systems have been used to code the Two Bags Task; however, they are generally similar and involve global ratings (on a 4- or 7-point scale) of salient aspects of parent and child behavior. Because this coding system is global, each case can be coded in real-time on one pass through the videotape. Total coding time would be about 12–14 minutes for the Two Bags Task, whereas the average coding time for the NCATS during the 9-month national study was approximately 17 minutes. Coding time per tape was an important consideration for a large-scale study, which could involve coding up to 10,000 videotapes.

In addition, the Two Bags Task is more straightforward and efficient to administer than the NCATS. The parent is handed the two bags and asked to play with the child for 10 minutes. The NCATS, on the other hand, has complicated and sometimes ambiguous instructions in which the parent must review a list of age-appropriate activities and select the first activity that the child cannot do. Task selection is verified by asking the mother if there is another activity after the selected task that the child can do. If there is, then the next task after that is selected. Task selection for the NCATS in the ECLS-B was difficult because often parents selected a task that was too young so that they could be assured that the child would be able to perform it on the videotape, or too old because the parent wanted the child’s precocious abilities on videotape. For these administrative reasons, the Two Bags Task would be less burdensome in the field and obtain more reliable information because all children receive the same tasks.

## 6.3 Two Bags Task Protocol for In-Home Administration

The Two Bags Task is a videotaped interaction. Therefore, interviewers administering the Two Bags Task during the home visit used a handheld video camera to film the parent and child engaging in the two activities that comprise the Two Bags Task. During the national training, interviewers were taught to administer and to videotape the Two Bags Task. The training included extensive practice, emphasizing good filming techniques and skillful use of the camera in conjunction with faithful administration of the Two Bags Task.

The Two Bags Task administration during the home visit was standardized to ensure that all interviewers administered the task in the same way to all parent-child pairs. To ensure this standardization, step-by-step Two Bags Task administration instructions were included in the Child Activity Booklet in a separate tabbed section for the Two Bags Task. These instructions included a verbatim script that was read to the parent. Interviewers also asked parents whether or not they had previously read *Good Night, Gorilla* to their child, and if so, how often. Interviewers were expected to record parents’ answers in check boxes on the administration pages in the Child Activity Booklet, making sure to record verbatim answers related to frequency. The interviewer also recorded the start time of the Two Bags Task and the language used by the parent when talking to the child.

In the case of twins, the interviewer administered the Two Bags Task separately for each twin, but recorded both on the same videotape and used the same activities. This introduced the problem of familiarity with the storybook as a confounding variable. It was possible that on the second reading of the storybook, the parent would alter the reading in some systematic way. Therefore, interviewers were instructed to counterbalance the administration of the Two Bags Task to twins. It would not be possible to impose a true counterbalanced design (in which, say, a random number generator was used to determine order of administration to all twin pairs in the ECLS-B before the 2-year data collection began), however, because this would have been too burdensome to field staff and probably not a realistic expectation. Therefore, field staff was instructed to administer the Two Bags Task to the first-born twin on odd-numbered days and to the second-born twin on even-numbered days. Field staff also recorded in the Child Activity Booklet which twin had been administered the Two Bags Task and in what order.

Unlike at 9 months, when a triadic NCATS involving the mother and both twins simultaneously was obtained after the mother completed the NCATS with each twin separately, there was no triadic Two Bags Task. For further information about triadic NCATS activities, 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).

After completion of the home visit, the field representative then sent the Two Bags Task videotape and the Child Activity Booklet, along with other data collection materials, to Westat's home office for receipting and coding by expert coders.

## 6.4 Two Bags Task Field Staff Training, Trainer Training, and Coder Training

Three different types of training were required for the Two Bags Task. The first was field staff training. Field staff was trained to obtain high-quality videotapes and to administer the Two Bags Task to the parent and child according to standardized procedures. Second was the trainer training. Home office staff targeted to train coders on the Two Bags Task coding system attended a training session at Columbia University Teachers College. The third was coder training held at the home office. Two Bags Task coders participated in extensive training to ensure reliability[1] of coding comparable to Teachers College standards.

[1] Reliability in this case refers to inter-rater reliability, which is the degree to which different raters or observers give consistent ratings to the same observed behaviors from the same videotapes.

Each of these trainings is described in more detail in the following sections. These descriptions are followed by a summary of quality control procedures that were followed to prevent coder drift from the standards as the year of data collection progressed. The final section summarizes how the Two Bags Task performed in the ECLS-B and presents information about inter-lab agreement between the trainers compared with the coding supervisor at Teachers College and intra-lab reliability between the coders and the reliability consensus coding by the Westat coding supervisor and assistant trainers.

### 6.4.1 Field Staff Training

For the 2-year data collection, some field staff were returnees from the 9-month data collection and some were new to the ECLS-B. Returning field staff already knew how to operate the videocamera used to tape the parent-child interactions. To enable new field staff to become familiar with the videocamera prior to training and thereby reduce the amount of time required during training, an 8 mm videocamera and an 8 mm cassette, together with an accompanying manual, were sent to each new trainee prior to the national training in Los Angeles. The trainees were instructed to follow the instructions in the manual and to practice using the videocamera at their convenience before coming to training. In addition, the field representative manual provided to all trainees included detailed instructions on videotaping and administering the Two Bags Task. Interviewers were able to refer to this manual during the field period as needed.

By the time of the national training, all trainees were familiar with the operation of the videocamera. This enabled attention to be focused directly on the correct administration of the Two Bags Task procedures at the national training. Trainees were instructed to follow the Two Bags Task administration steps verbatim as presented in the Child Activity Booklet. They then administered the Two Bags Task to each other in sets of three, in alternating turns, one playing the role of the interviewer, one the parent, and one the child.

National training did include emphasis on proper videotaping techniques to obtain a high-quality videotape of the Two Bags Task interaction. A high-quality videotape was critical to successful Two Bags Task coding. Therefore, trainees received hands-on practice and extensive feedback about their videotaping. This was done during the sessions involving direct instruction and also during the live-practice session when training staff circulated through the rooms and watched over the shoulders of field staff as they videotaped their partners administering the Bayley Short Form–Research Edition (BSF-R). Westat staff members reviewing the videotape to score the BSF-R administration also reviewed the quality of the videotaping. Any videotape that was not of sufficient quality (e.g., audio level too low, lighting level too low, camera faced toward a window so that the dyad was seen only in silhouette, etc.) was noted, and the videotaper was required to attend a help session and/or demonstrate good videotaping skills to her or his lead trainer. In this way, all field staff who had trouble producing a high-quality videotape received intervention and retraining before going into the field.

In addition, videotapes from each field staff member were quality reviewed by the Two Bags Task coding staff on an ongoing basis as they were received at Westat. Feedback on videotape quality was given to all field staff within about a week of receipt of the videotapes. For further information about quality control procedures, please see section 6.5.

### 6.4.2 Trainer Training

The first task of training the trainers was to have them trained on the coding system so that they, in turn, could train the individuals who would actually be coding the videotapes. The trainer training was done by a graduate student working in the laboratory of Dr. Jeanne Brooks-Gunn at Teachers College, together with Christy Brady-Smith, the first author of their coding manual (Brady-Smith et al. 1999). This individual had been the reliability coder for the Three Bags Task used in the Early Head Start Research and Evaluation Study, thereby ensuring that the Westat Two Bags Task trainers would be trained to the same standards as those used in the Early Head Start study and that results would be comparable to that study.

In May 2003, four Westat staff members from the Child and Family Studies area attended the specialized training at Teachers College. The training took place over the span of 3 days and was conducted by the lead coder for the Early Head Start Research and Evaluation Study. Instruction over the first 2 days consisted of a review of the rating scales interspersed with videoclips of examples of the types of behaviors in the rating scales. The third day consisted of reliability coding of seven videotapes. To pass the training, the Westat staff members had to code each videotape and score within 90 percent agreement with the reliability coding.

The procedure followed by the Early Head Start Research and Evaluation Project had been that, once an individual had passed the training, inter-lab reliabilities were provided for a maximum of the first 30 videotapes after training, or until the coder could sustain 90 percent agreement with the reliability coding for 5 consecutive tapes, whichever came first. The same procedure was followed for the Westat trainers. After completion of training, each trainer coded reliability tapes provided by Teachers College. Tapes were sent in batches of 5 to Westat. The trainers coded the tapes and sent their coding sheets to CUTC in batches. The actual scoring of the reliability testing for certification was done by the lead Three Bags Task coder at Teachers College. Westat's trainers established reliability on all the Two Bags Task rating scales quickly and only required an average of 12 reliability tapes before becoming reliable at 90 percent agreement for all the rating scales.

One staff member, who had been the coding supervisor for the NCATS during the 9-month data collection, was designated to be the coding supervisor for the Two Bags Task and lead trainer for the training of coders. A second individual co-led this training and was designated to be the back-up for the supervisor. Together, the coding supervisor and this assistant consensus coded[2] all reliability tapes that were used to establish intra-lab reliability between the coders and the reliability videotapes. A third individual, a member of the Child Development Team, was also called upon to resolve coding questions that arose during the course of the year and also coded incoming videotapes as necessary, depending on the work load. The fourth individual served as the liaison between the Child and Family Studies area and the Two Bags Task coding workshop. The coding supervisor, her assistants, and the liaison were able to establish, maintain, and share with coders a Two Bags Task coding knowledge base that contributed to the maintenance of coding reliability across the data collection year.

### 6.4.3 Coder Training

All six NCATS coders from the 9-month data collection who were still on staff at the beginning of the 2-year data collection were retained to code Two Bags Task videotapes. These coders had already demonstrated their coding competence with weekly NCATS reliabilities above the 85 percent agreement criterion. An additional coder, recruited from within Westat, completed the observation skills test that had been used in the recruitment of NCATS coders and passed it at greater than 90 percent. The analyst is referred to the *ECLS-B Methodology Report for the Nine-Month Data Collection, Volume 1:*

[2] Consensus coding in this context, refers to the process where two (or more) individuals each code a videotape independently and then compare their ratings item-by-item. If there are any discrepancies between the ratings, the discrepancies are resolved by discussion. The result is a final set of ratings that can be used as a standard against which to compare the ratings of other coders.

*Psychometric Characteristics* (NCES 2005–100) (Andreassen and Fletcher 2005) for further information about this observation skills test.

The training for the seven coder trainees was held about 3 weeks after the trainer training. During that interval, the trainers had established their ongoing reliability on the rating scales, prepared the coder training materials tailored to the specific needs of the Two Bags Task in the ECLS-B and obtained videotapes for training and reliability purposes. The coder training took a full 5 days. The first 3 days were devoted to an introduction to the nine Two Bags Task rating scales, interspersed with videoclips of mother-child interactions to demonstrate the target interaction behaviors, with additional coding practice devoted to one rating scale at a time. The fourth day provided practice in coding Two Bags Task videotapes on all the rating scales simultaneously. The culmination was on the fifth day, when trainees completed seven reliability videotapes. All trainees passed these reliability tapes at the required 90 percent agreement or greater. Following training, the supervisor and her assistants provided coding support to the new coders on an as-needed basis. In addition, if a coder encountered a videotape that was difficult to code, it was brought to the attention of the supervisor who conducted a weekly “brown-bag” coding review session to discuss coding issues and difficulties that may have arisen during the week.

Initially, only English videotapes were coded because none of the coders were fluent in any other major language, such as Spanish or Chinese. Therefore, all foreign language videotapes were put aside. After a sufficient number of foreign language videotapes had been assembled, new coders fluent in the required languages were added to the coding staff and two subsequent trainings were conducted on an as-needed basis. The second training for two additional coders, one fluent in Mandarin and one fluent in Spanish, occurred 4 months after the first training, and was led by the coding supervisor and assistant. A third training became necessary when the Spanish-speaking coder resigned, and a new one had to be hired. This training involved only this one Spanish-speaking trainee and the coding supervisor, who followed the same training script and procedures as those used for the other trainings.

## 6.5 Two Bags Coding Quality Control Procedures and Reliability

In keeping with procedures instituted for the 9-month NCATS coding, coders worked up to 4 hours a day, coding a maximum of 10 videotapes. This limitation was implemented to maintain reliability of coding and to prevent "coder drift."[3] Coding reliability begins to falter beyond that number. Initially, all coders worked up to 4 hours a day as coders and then spent the other 4 hours working on other ECLS-B activities, such as field staff payroll, locating respondents, receipting, and computer-assisted data entry (CADE).

Unlike the NCATS coding at 9 months, when inter-lab reliability was maintained between Westat and the NCAST coder at the University of Washington, inter-lab reliability between Westat and the Three Bags Task coder at Teachers College was not maintained beyond the initial training period as had been the case for the Early Head Start Research and Evaluation Project. The Three Bags Task coding staff at Teachers College did not see the need for, and did not have the resources to provide, ongoing inter-lab reliability. Therefore, reliability was maintained within the Westat coding workshop. Intra-lab reliability required that the coding supervisor select a random subsample of ECLS-B Two Bags Task videotapes. The supervisor and an assistant (trained at Teachers College) coded each selected tape independently and then resolved any discrepancies by consensus. When necessary, a third individual (also trained at Teachers College) was brought in to help resolve any particularly difficult coding issues. These selected videotapes ("reliability videotapes") were then used to establish the reliability of the coding staff.

Two Bags Task coders were then required to code one reliability tape per week, selected at random. Coders were required to code the weekly reliability videotape with a minimum of 85 percent agreement with the consensus reliability coding. If a coder slipped below 85 percent agreement on a weekly reliability videotape, that coder then immediately coded a second reliability videotape. Had there been a case where the second reliability videotape was also below 85 percent agreement, the coder would have been told to cease coding any videotapes from the ECLS-B and would have received remedial training on the identified coding problems. In practice, however, no coder ever slipped below 85 percent agreement on more than one reliability videotape. Agreement rates between the coders and the reliability coding were quite high for the entire coding year and are summarized in table 6-1.

[3] Coder drift refers to the change in how information is coded over time by an individual. The coder is said to "drift" from the standard due to such factors as fatigue, forgetting of the rules, failing to detect target information, among others.

Table 6-1. Average reliability (percent agreement) for subscales of the Two Bags Task for the ECLS-B 2-year data collection: 2003–04

| Two Bags Task scale | Mean percentage agreement |
|---|---|
| Overall agreement for parent rating scales: | 96.5 |
| Parent rating scales | |
| Sensitivity | 97.0 |
| Intrusiveness | 98.0 |
| Positive regard | 93.0 |
| Cognitive stimulation | 94.0 |
| Negative regard | 98.0 |
| Detachment | 99.0 |
| Overall agreement for child rating scales | 94.7 |
| Child rating scales | |
| Engagement | 94.0 |
| Sustained attention | 93.0 |
| Negativity | 97.0 |

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.

## 6.6 Two Bags Performance in the 2-Year National Data Collection

In the Early Head Start Research and Evaluation Project, three of the parent scales were intercorrelated: parental sensitivity, parental stimulation of cognitive development, and the parental positive regard scale. These three variables were combined by Early Head Start to create a Supportiveness composite at 2 years by simply obtaining the mean of the three scales (i.e., the sum of the scores for Parental Sensitivity, Parental Cognitive Stimulation, and Parental Positive Regard divided by 3). For the convenience of researchers, a composite of the average of these three variables also was created for the ECLS-B 2-year data collection. This composite is X2TBSPPT.

Table 6-2 presents descriptive statistics for the Two Bags Task variables for the sample as a whole.

Table 6-2. Weighted means and standard deviations for the Two Bags Task rating scales in the ECLS-B 2-year data collection: 2003–04

| Two Bags Task scale | Number | Range | Weighted mean | Standard deviation |
|---|---|---|---|---|
| Parent rating scales | | | | |
| Sensitivity (C2SENSTV) | 7,450 | 1-7 | 4.77 | 0.95 |
| Intrusiveness (C2NTRUSV) | 7,450 | 1-7 | 1.80 | 0.54 |
| Positive regard (C2POSRGD) | 7,450 | 1-7 | 4.26 | 1.03 |
| Cognitive stimulation (C2COGDEV) | 7,450 | 1-7 | 4.12 | 1.08 |
| Negative regard (C2NEGRGD) | 7,450 | 1-7 | 1.10 | 0.44 |
| Detachment (C2DETACH) | 7,450 | 1-7 | 1.05 | 0.32 |
| Supportiveness (X2TBSPPT) | 5,600 | 1-7 | 4.43 | 0.86 |
| Child rating scales | | | | |
| Engagement (C2ENGPRT) | 7,450 | 1-7 | 4.56 | 1.14 |
| Sustained attention (C2STNATT) | 7,450 | 1-7 | 4.47 | 1.15 |
| Negativity (C2NEGPRT) | 7,450 | 1-7 | 1.36 | 0.76 |

NOTE: The child weight W2C0 was used to produce these statistics. The variable name of the scale is in parentheses. Sample sizes have been rounded to the nearest 50.
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.

Although the scores obtained are considered rating scales, they could also be conceptualized as items. Therefore, Cronbach’s alpha was also calculated to investigate whether the scales have a conceptual coherence that would make it feasible to scale them into a single scale. This was done only for the parent scales and did not include the composite X2TBSPPT. Cronbach’s alpha was not calculated for the child scales because there are only three reducing the accuracy of Cronbach’s alpha. For all the scales, a value of 1 indicates a low score for that dimension and a 7 indicates a high score. Therefore, it was not necessary to reverse code any of the scores. Cronbach’s alpha for the parent scale was 0.73, indicating that these items have adequate coherence, and the analyst may want to consider creating a single scale comprised of all six parent scales. That said, however, the analyst is cautioned that the rating scales were not designed with that intent and that using them in this way may reduce their usefulness. Instead, the analyst may want to investigate the factor structure of these scores and conduct a factor analysis.

## 6.7 Correlations of 2-Year Two Bags Task and 9-Month NCATS Scale Scores

Although the Two Bags Task rating scales and the NCATS scale used at 9 months have their differences, the constructs they measure do share some similarities. The NCATS parent scales include items that assess parental sensitivity and responsiveness to the child’s distress, as well as parent fostering of the child's socioemotional and cognitive growth. The NCATS child scales include items that assess the child's responsiveness to the parent and the child's clarity of communication to the parent. Therefore, both scales measure maternal sensitivity and engagement with the child, as well as the child's engagement with the parent. The Two Bags Task also shares some characteristics with the child's responsivity to the parent on the NCATS scale. Therefore, meaningful correlations should be found between the two scales. Table 6-3 summarizes the correlations between the 2-year Two Bags Task rating scales and the 9-month NCATS scales. To obtain these correlations, all cases with missing data were omitted and the child weight, W2C0, was applied.

Table 6-3. Correlation (*r*) of 2-year Two Bags Task rating scales with 9-month NCATS total scale, total parent scale and total child scale, ECLS-B 9-month and 2-year data collections: 2001–02 and 2003–04

| | 9-month NCATS scales | | |
|---|---|---|---|
| Two Bags Task scale | Total scale correlation (*r*) | Total parent scale correlation (*r*) | Total child scale correlation (*r*) |
| Parent rating scales | | | |
| Sensitivity (C2SENSTV) | .19* | .21* | .05* |
| Intrusiveness (C2NTRUSV) | -.09* | -.11* | -.00 |
| Positive regard (C2POSRGD) | .18* | .21* | .05* |
| Cognitive stimulation (C2COGDEV) | .17* | .19* | .06* |
| Negative regard (C2NEGRGD) | -.11* | -.13* | -.02 |
| Detachment (C2DETACH) | -.05* | -.05* | -.04* |
| Supportiveness (X2TBSPPT) | .22* | .24* | .06* |
| Child rating scales | | | |
| Engagement (C2ENGPRT) | .14* | .15* | .05* |
| Sustained attention (C2STNATT) | .10* | .10* | .03* |
| Negativity (C2NEGPRT) | -.07* | -.09* | .01 |

* *p* < .05

NOTE: The child weight W2C0 was used to obtain these data. *n* = 4,900 (rounded to the nearest 50).

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.

## 6.8 Two Bags Task Measures in the 2-Year Data Collection

Table 6-4 summarizes the Two Bags Task parent scales that assess characteristics of the primary caregiver's interaction with the child and table 6-5 summarizes the children's scales. Both tables present the means and standard deviations for the total sample and by key demographic groups.

Table 6-4. Weighted means (and standard deviations) of the Two Bags Task parent scales by key demographic variables, 2-year data collection: 2003–04

| | | Two Bags Task parent scale variables, weighted means | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Supportiveness composite (X2TBSPPT) | Sensitivity (C2SENSTV) | Intrusiveness (C2NTRUSV) | Positive regard (C2POSRGD) | Cognitive stimulation (C2COGDEV) | Negative regard (C2NEGRGD) | Detachment (C2DETACH) |
| Total sample | 7,450 | 4.38 (0.88) | 4.79 (1.62) | 1.19 (1.14) | 4.29 (1.83) | 4.13 (1.64) | 1.12 (0.44) | 1.09 (1.90) |
| Child’s race/ethnicity[1] | | | | | | | | |
| White | 3,250 | 4.59 (0.82) | 5.00 (0.86) | 1.12 (0.61) | 4.50 (2.24) | 4.35 (1.99) | 1.07 (0.33) | 1.09 (2.45) |
| Black | 1,200 | 4.05 (0.83) | 4.37 (0.95) | 1.41 (0.82) | 3.90 (1.08) | 3.86 (0.96) | 1.32 (0.77) | 1.07 (0.40) |
| Hispanic, race specified | 1,000 | 4.20 (0.90) | 4.71 (3.48) | 1.27 (2.47) | 4.09 (1.04) | 3.92 (1.02) | 1.10 (0.40) | 1.07 (0.32) |
| Hispanic, no race specified | 450 | 3.95 (0.90) | 4.31 (0.99) | 1.16 (0.44) | 3.85 (0.98) | 3.69 (1.02) | 1.10 (0.36) | 1.13 (0.50) |
| Asian | 700 | 4.18 (0.87) | 4.52 (0.94) | 1.24 (0.60) | 4.13 (1.03) | 3.90 (1.04) | 1.15 (0.50) | 1.21 (3.11) |
| Native Hawaiian/Pacific Islander | 50 | 4.25 (0.63) | 4.52 (0.66) | 1.27 (0.54) | 4.57 (1.04) | 3.66 (0.71) | 1.10 (0.30) | 1.04 (0.27) |
| American Indian/Alaska Native | 200 | 3.93 (0.78) | 4.53 (0.85) | 1.08 (0.33) | 3.76 (1.03) | 3.50 (1.01) | 1.09 (0.36) | 1.07 (0.28) |
| More than 1 race | 600 | 4.46 (0.84) | 4.87 (0.94) | 1.13 (0.50) | 4.40 (1.07) | 4.10 (1.04) | 1.09 (0.33) | 1.04 (0.29) |
| Poverty status | | | | | | | | |
| Below poverty threshold | 1,600 | 3.94 (0.85) | 4.40 (3.06) | 1.37 (2.22) | 3.82 (1.07) | 3.68 (0.97) | 1.23 (0.64) | 1.10 (0.45) |
| At or above poverty threshold | 5,850 | 4.50 (0.85) | 4.90 (0.92) | 1.14 (0.59) | 4.41 (1.96) | 4.25 (1.76) | 1.08 (0.37) | 1.09 (2.12) |
| Child’s sex | | | | | | | | |
| Male | 3,800 | 4.34 (0.88) | 4.77 (2.07) | 1.22 (1.52) | 4.23 (1.04) | 4.06 (1.08) | 1.14 (0.49) | 1.12 (2.55) |
| Female | 3,650 | 4.43 (0.88) | 4.82 (0.95) | 1.16 (0.49) | 4.34 (2.39) | 4.21 (2.07) | 1.09 (0.38) | 1.06 (0.77) |

See notes at end of table.

Table 6-4. Weighted means (and standard deviations) of the Two Bags Task parent scales by key demographic variables, 2-year data collection: 2003–04—Continued

| | | Two Bags Task parent scale variables, weighted means | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Supportiveness composite (X2TBSPPT) | Sensitivity (C2SENSTV) | Intrusiveness (C2NTRUSV) | Positive regard (C2POSRGD) | Cognitive stimulation (C2COGDEV) | Negative regard (C2NEGRGD) | Detachment (C2DETACH) |
| Birth weight | | | | | | | | |
| Normal | 5,500 | 4.39 (0.88) | 4.81 (1.67) | 1.17 (0.64) | 4.29 (1.88) | 4.15 (1.68) | 1.11 (0.44) | 1.09 (1.97) |
| Moderately low | 1,150 | 4.29 (0.84) | 4.64 (0.92) | 1.40 (3.86) | 4.20 (1.04) | 4.02 (1.01) | 1.17 (0.51) | 1.06 (0.40) |
| Very low | 800 | 4.19 (0.86) | 4.53 (0.95) | 1.32 (0.74) | 4.18 (1.05) | 3.86 (1.03) | 1.18 (0.53) | 1.06 (0.37) |
| | | | | | | | | |
| Child’s age at assessment | | | | | | | | |
| 21 months and under | # | 4.66 (0.67) | 4.58 (0.49) | 1.00 (0.00) | 4.76 (0.61) | 4.65 (1.12) | 1.00 (0.00) | 1.00 (0.00) |
| 22–23 months | 750 | 4.24 (0.89) | 4.58 (0.95) | 1.23 (0.61) | 4.15 (1.03) | 3.98 (1.06) | 1.16 (0.50) | 1.07 (0.38) |
| 24–25 months | 5,750 | 4.40 (0.88) | 4.83 (1.77) | 1.17 (0.65) | 4.30 (2.00) | 4.15 (1.78) | 1.11 (0.43) | 1.10 (2.15) |
| 26–27 months | 750 | 4.39 (0.90) | 4.77 (0.97) | 1.27 (3.16) | 4.27 (1.06) | 4.13 (1.07) | 1.11 (0.45) | 1.07 (0.40) |
| 28 months and over | 150 | 4.46 (0.78) | 4.84 (0.95) | 1.14 (0.41) | 4.38 (0.88) | 4.17 (0.91) | 1.11 (0.34) | 1.02 (0.16) |
| | | | | | | | | |
| Mother’s race/ethnicity[1] | | | | | | | | |
| White | 3,550 | 4.59 (0.82) | 5.00 (0.87) | 1.12 (0.61) | 4.49 (2.19) | 4.33 (1.95) | 1.07 (0.33) | 1.09 (2.37) |
| Black | 1,200 | 4.05 (0.83) | 4.38 (0.96) | 1.41 (0.82) | 3.91 (1.08) | 3.87 (0.95) | 1.33 (0.80) | 1.07 (0.40) |
| Hispanic, race specified | 1,200 | 4.07 (0.91) | 4.51 (3.05) | 1.24 (2.14) | 3.99 (1.02) | 3.80 (1.02) | 1.09 (0.35) | 1.10 (0.42) |
| Hispanic, no race specified | 50 | 3.92 (0.96) | 4.39 (1.02) | 1.06 (0.36) | 3.65 (0.88) | 3.73 (1.17) | 1.04 (0.23) | 1.14 (0.45) |
| Asian | 850 | 4.22 (0.85) | 4.57 (0.93) | 1.21 (0.56) | 4.16 (1.02) | 3.93 (1.04) | 1.13 (0.47) | 1.19 (2.81) |
| Native Hawaiian/Pacific Islander | 50 | 4.36 (0.79) | 4.69 (0.92) | 1.32 (0.62) | 4.38 (1.05) | 4.00 (1.03) | 1.16 (0.37) | 1.04 (0.28) |
| American Indian/Alaska Native | 250 | 4.09 (0.80) | 4.69 (0.87) | 1.08 (0.36) | 3.96 (1.05) | 3.61 (1.00) | 1.07 (0.28) | 1.05 (0.26) |
| More than 1 race | 200 | 4.40 (0.86) | 4.83 (0.98) | 1.14 (0.54) | 4.30 (1.03) | 4.07 (1.02) | 1.13 (0.39) | 1.00 (0.07) |

See notes at end of table.

Table 6-4. Weighted means (and standard deviations) of the Two Bags Task parent scales by key demographic variables, 2-year data collection: 2003–04—Continued

| | | Two Bags Task parent scale variables, weighted means | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Number | Supportiveness composite (X2TBSPPT) | Sensitivity (C2SENSTV) | Intrusiveness (C2NTRUSV) | Positive regard (C2POSRGD) | Cognitive stimulation (C2COGDEV) | Negative regard (C2NEGRGD) | Detachment (C2DETACH) |
| Mother’s age (in years) | | | | | | | | |
| 19 and under | 250 | 3.85 (0.94) | 4.24 (1.03) | 1.38 (0.80) | 3.71 (1.18) | 3.60 (0.98) | 1.40 (0.93) | 1.10 (0.45) |
| 20–29 | 3,300 | 4.23 (0.87) | 4.63 (0.96) | 1.23 (1.50) | 4.10 (1.04) | 3.95 (1.04) | 1.14 (0.49) | 1.13 (2.66) |
| 30–39 | 3,250 | 4.56 (0.85) | 4.95 (0.91) | 1.13 (0.46) | 4.50 (2.50) | 4.31 (1.08) | 1.07 (0.33) | 1.05 (0.79) |
| 40 and over | 550 | 4.59 (0.85) | 5.25 (5.24) | 1.14 (1.39) | 4.48 (0.97) | 4.62 (5.01) | 1.07 (0.31) | 1.04 (0.31) |
| | | | | | | | | |
| Mother’s education | | | | | | | | |
| 8th grade and under | 300 | 3.67 (0.81) | 4.01 (0.95) | 1.20 (0.52) | 3.61 (0.93) | 3.40 (0.84) | 1.11 (0.43) | 1.22 (0.63) |
| 9–12th grades | 1,400 | 4.00 (0.87) | 4.47 (3.07) | 1.30 (0.73) | 3.89 (1.07) | 3.74 (1.00) | 1.21 (0.61) | 1.26 (4.02) |
| High school diploma | 1,600 | 4.23 (0.83) | 4.62 (0.94) | 1.27 (2.11) | 4.12 (1.02) | 3.94 (0.98) | 1.15 (0.51) | 1.04 (0.26) |
| Vocational/technical | 150 | 4.41 (0.77) | 4.81 (0.90) | 1.13 (0.44) | 4.28 (0.84) | 4.12 (1.00) | 1.04 (0.20) | 1.05 (0.29) |
| Some college | 1,850 | 4.53 (0.82) | 4.94 (0.85) | 1.14 (0.81) | 4.52 (3.18) | 4.23 (1.07) | 1.09 (0.36) | 1.03 (0.19) |
| Bachelor’s degree | 1,250 | 4.78 (0.77) | 5.16 (0.80) | 1.08 (0.35) | 4.60 (0.91) | 4.68 (3.25) | 1.05 (0.27) | 1.04 (1.23) |
| Graduate school (no degree) | 150 | 4.86 (0.63) | 5.27 (0.69) | 1.05 (0.28) | 4.71 (0.78) | 4.59 (1.04) | 1.05 (0.22) | 1.00 (0.06) |
| Master’s degree | 500 | 4.86 (0.72) | 5.27 (0.80) | 1.08 (0.38) | 4.69 (0.84) | 4.63 (1.06) | 1.03 (0.23) | 1.02 (0.17) |
| Doctoral/professional degree | 200 | 4.89 (0.77) | 5.25 (0.78) | 1.01 (0.14) | 4.68 (0.82) | 4.73 (1.00) | 1.02 (0.16) | 1.03 (0.34) |

# Rounds to zero.

[1] Race categories exclude Hispanic origin unless specified.

NOTE: Results were obtained by applying the sampling child weight W2C0. The variable names of the parent scales are in parentheses in column headings. Standard deviations appear in parentheses in the table columns. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

Table 6-5. Weighted means (and standard deviations) of the Two Bags Task child scales by key demographic variables, 2-year data collection: 2003–04

| | | Two Bags Task child scale variables, weighted means | | |
|---|---|---|---|---|
| Characteristic | Number | Engagement (C2ENGPRT) | Sustained attention (C2STNATT) | Negative regard (C2NEGPRT) |
| Total sample | 7,450 | 4.57 (1.30) | 4.51 (2.18) | 1.39 (1.95) |
| | | | | |
| Child's race/ethnicity[1] | | | | |
| White | 3,250 | 4.77 (1.23) | 4.72 (2.68) | 1.35 (2.15) |
| Black | 1,200 | 4.32 (1.08) | 4.25 (1.06) | 1.60 (2.77) |
| Hispanic, race specified | 1,000 | 4.30 (1.21) | 4.24 (1.18) | 1.40 (0.84) |
| Hispanic, no race specified | 450 | 4.14 (1.93) | 4.06 (1.86) | 1.34 (0.71) |
| Asian | 700 | 4.32 (1.07) | 4.43 (1.09) | 1.38 (0.77) |
| Native Hawaiian/Pacific Islander | 50 | 4.38 (0.98) | 4.22 (0.93) | 1.27 (0.64) |
| American Indian/Alaska Native | 200 | 4.21 (0.98) | 3.84 (1.20) | 1.48 (0.88) |
| More than 1 race | 600 | 4.71 (1.09) | 4.56 (1.07) | 1.32 (0.73) |
| | | | | |
| Poverty status | | | | |
| Below poverty threshold | 1,600 | 4.15 (1.48) | 4.09 (1.43) | 1.49 (0.89) |
| At or above poverty threshold | 5,850 | 4.67 (1.22) | 4.61 (2.33) | 1.37 (2.14) |
| | | | | |
| Child's sex | | | | |
| Male | 3,800 | 4.44 (1.30) | 4.35 (2.06) | 1.46 (2.26) |
| Female | 3,650 | 4.70 (1.28) | 4.67 (2.29) | 1.32 (1.54) |
| | | | | |
| Birth weight | | | | |
| Normal | 5,500 | 4.59 (1.31) | 4.52 (2.25) | 1.39 (2.10) |
| Moderately low | 1,150 | 4.37 (1.09) | 4.31 (1.10) | 1.44 (0.84) |
| Very low | 800 | 4.07 (1.14) | 4.00 (1.05) | 1.51 (0.81) |
| | | | | |
| Child's age at assessment | | | | |
| 21 months and under | # | 4.26 (0.47) | 4.14 (0.66) | 1.34 (0.48) |
| 22–23 months | 750 | 4.30 (1.71) | 4.27 (1.71) | 1.38 (0.76) |
| 24–25 months | 5,750 | 4.59 (1.24) | 4.52 (2.35) | 1.41 (2.17) |
| 26–27 months | 750 | 4.63 (1.17) | 4.54 (1.14) | 1.33 (0.73) |
| 28 months and over | 150 | 4.83 (1.11) | 4.85 (1.16) | 1.13 (0.35) |

See notes at end of table.

Table 6-5. Weighted means (and standard deviations) of the Two Bags Task child scales by key demographic variables, 2-year data collection: 2003–04—Continued

| | | Two Bags Task child scale variables, weighted means | | |
|---|---|---|---|---|
| Characteristic | Number | Engagement (C2ENGPRT) | Sustained attention (C2STNATT) | Negative regard (C2NEGPRT) |
| Mother's race/ethnicity[1] | | | | |
| White | 3,550 | 4.76 (1.22) | 4.70 (2.61) | 1.35 (2.10) |
| Black | 1,200 | 4.33 (1.09) | 4.26 (1.05) | 1.61 (2.76) |
| Hispanic, race specified | 1,200 | 4.21 (1.45) | 4.16 (1.50) | 1.38 (0.30) |
| Hispanic, no race specified | 50 | 4.02 (1.47) | 4.12 (1.37) | 1.18 (0.47) |
| Asian | 850 | 4.33 (1.06) | 4.43 (1.08) | 1.34 (0.72) |
| Native Hawaiian/Pacific Islander | 50 | 4.53 (0.89) | 4.11 (0.98) | 1.25 (0.61) |
| American Indian/Alaska Native | 250 | 4.35 (1.01) | 4.07 (1.20) | 1.40 (0.84) |
| More than 1 race | 200 | 4.71 (1.12) | 4.50 (1.13) | 1.39 (0.82) |
| Mother's age (in years) | | | | |
| 19 and under | 250 | 4.29 (1.14) | 4.17 (1.13) | 1.44 (0.91) |
| 20–29 | 3,300 | 4.47 (1.32) | 4.42 (2.34) | 1.43 (1.62) |
| 30–39 | 3,250 | 4.70 (1.30) | 4.62 (2.19) | 1.35 (2.41) |
| 40 and over | 550 | 4.57 (1.08) | 4.51 (1.11) | 1.31 (0.60) |
| Mother's education | | | | |
| 8th grade and under | 300 | 3.78 (1.19) | 3.78 (1.11) | 1.42 (0.77) |
| 9–12th grades | 1,400 | 4.26 (1.16) | 4.30 (3.15) | 1.47 (0.86) |
| High school diploma | 1,600 | 4.46 (1.10) | 4.36 (1.10) | 1.39 (0.80) |
| Vocational/technical | 150 | 4.63 (3.17) | 4.56 (3.15) | 1.23 (0.63) |
| Some college | 1,850 | 4.68 (1.08) | 4.57 (1.11) | 1.39 (2.07) |
| Bachelor's degree | 1,250 | 4.89 (1.54) | 4.85 (3.25) | 1.40 (3.81) |
| Graduate school (no degree) | 150 | 4.96 (1.05) | 4.74 (1.25) | 1.32 (0.83) |
| Master's degree | 500 | 5.03 (1.01) | 4.89 (1.11) | 1.20 (0.67) |
| Doctoral/professional degree | 200 | 4.97 (0.94) | 4.99 (1.08) | 1.21 (0.51) |

# Rounds to zero.

[1] Race categories exclude Hispanic origin unless specified.

NOTE: Results were obtained by applying the sampling child weight W2C0. Standard deviations appear in parentheses. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

# 7. PHYSICAL MEASUREMENTS

This chapter presents a brief overview of the physical measurements obtained at 2 years. Training and quality control procedures are summarized, and correlational evidence is presented of the reliability of the measurements obtained during the 2-year data collection.

Physical growth measurements, as well as early motor development and early health care, are important constructs that were assessed in this study and are thought to be important factors contributing to school readiness. Children grow rapidly from birth through the early childhood years, requiring periodic key growth measurements. These periodic measurements, including child height, weight, middle upper arm circumference (MUAC), and head circumference for children born at very low birth weight (1,500 grams or less), were obtained because they are generally recognized as being accurate indicators of children’s nutrition, health status, physical development, and well-being.

## 7.1 Procedural Differences at 2 Years Compared to 9 Months

In the 9-month data collection, length, weight, and MUAC were obtained for all children. In addition, head circumference was obtained for children born at very low birth weight. To measure child length at 9 months, a measure mat was used because children this age cannot stand independently to be measured for height. With the child recumbent on the measure mat, a foot plate was placed at the soles of the child’s feet and the correct measurement of length was read from the markings on the measure mat. To measure child weight at 9 months, the mother first stepped on a SECA weight scale and her weight was recorded. With the mother remaining standing on the scale, the interviewer tapped a button on the scale to reset it and then handed the child to the mother. The scale then automatically calculated the child’s weight by subtracting the mother’s weight from the combined weight of mother and child. To measure MUAC, the child sat in the mother’s lap and the interviewer measured the length of the child’s upper arm and found the midpoint. To obtain the circumference of the upper arm, the interviewer looped a measuring tape around the child’s upper arm and tightened it at the midpoint. To obtain head circumference, with the child sitting in the mother’s lap, the interviewer looped the retractable tape measure around the child’s head, just above the brow and around the largest diameter in back. Head circumference was read at a point midway between the eyes just above the brow.

At 2 years, the same physical measurements were obtained, although the procedures varied a bit due to the child's more advanced physical development. Procedures for obtaining these measurements were adapted from the protocol for the National Health and Nutrition Examination Survey (NHANES), a major health and nutrition survey. In keeping with this protocol and with standard 9-month practice, all physical measurements were obtained twice.

At 2 years, child height rather than child length was measured. Because children at this age are able to follow basic instructions and to stand independently, a stadiometer, the Model 214 Road Rod by SECA, was used to obtain child height. With the child standing erect at the base of the stadiometer and with the child's head in correct position, a crown piece was lowered down the stadiometer ruler and child height was obtained in centimeters and recorded in the Child Activity Booklet. Child weight was obtained by instructing the child to stand independently on the SECA scale, and the measurement was recorded in the Child Activity Booklet. In addition, child height and weight were used to calculate the child's body mass index (BMI), based on a Centers for Disease Control and Prevention (CDC) formula available at its website, http://www.cdc.gov/nccdphp/dnpa/bmi/calc-bmi.htm. MUAC and head circumference were obtained using the same procedures as at 9 months, although it was no longer necessary for the child to sit on the mother's lap.

## 7.2 Two-Year Physical Measurement Variables on the Data File

Each physical measurement was obtained up to three times for each measure. If the first two measurements were within 5 percent of each other, the third measurement was not necessary. Interviewers were trained to obtain the first measurement and estimate 5 percent of the obtained value by figuring out what 10 percent would be and then dividing by 2. In those cases in which the second measurement was more than 5 percent different (either greater or smaller than the first measurement), a third measurement was obtained and recorded. All measurements were recorded on the appropriate record form in the Child Activity Booklet. For each measure, composite variables were created that indicate the average for each physical measurement. In addition, because child height at 2 years was obtained with the child standing erect, a composite for children's BMI could also be obtained. For more information about how these composites were created, please refer to the *User's Manual for the ECLS-B Longitudinal 9-Month–2-Year Data File and Electronic Codebook* (NCES 2006-046) (Nord et al. 2006). Table 7-1 summarizes the average weight (X2CHWGT), height (X2CHHGT), MUAC (X2CHMUAC), and BMI composites (X2CHBMI) for the key demographic groups, as well as average head circumference (X2CHCRFM) obtained for children born at very low birth weight. To obtain these statistics, all cases with missing data were omitted and the child weight W2C0 was applied.

## 7.3 Reliability of 2-Year Physical Measurements

Procedures established to ensure high data quality were implemented at training and continued throughout the year of data collection. First, interviewers were required to demonstrate correct procedures for obtaining the physical measurements. Second, once they had begun their home visits, a field supervisor or a Westat staff member accompanied the interviewer on a home visit and did a quality control review of procedures. Finally, when physical measurement data were entered at Westat, they were also routinely reviewed for errors. When systematic errors were found, the interviewer’s field supervisor was contacted and that person, in turn, contacted the interviewer and reviewed the physical measurements procedures. These procedures are reviewed in the following sections.

### 7.3.1 Training Procedures and Certification

During the national training, interviewers had the opportunity for hands-on practice in dyads to obtain the physical measurements and to demonstrate competence to the trainers. Interviewers were trained to obtain each physical measurement up to three times and to record each appropriately on the Physical Measurements record form. As described earlier, they were also trained to ascertain that the second measurement in a set was within 5 percent of the first one. If the difference between two measurements in a set was greater than 5 percent, interviewers were trained to obtain a third measurement and to record it in the appropriate space on the Physical Measurements record form.

Certification on the physical measurements was obtained during the physical measurements training session. Three measuring stations with the appropriate measurement equipment were set up in each room. In each training room, the trainees were divided into thirds for the certification exercise. The trainer, co-trainer, and training assistant went to one of the three stations and played the role of the focus child to be measured. The trainees were assigned to a station and instructed to collect the physical measurements of the training staff person twice, the same as in the field.

Table 7-1. Children's average physical measurements and standard deviations for total sample and by key demographic variables, 2-year data collection: 2003–04

| | Children's average physical measurements | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Weight | | | Height | | | Middle upper arm circumference | | | Head circumference[1] | | | Body mass index | | |
| Characteristic | Number | Mean (kg) | *SD* (kg) | Number | Mean (kg) | *SD* (kg) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) | Number | Mean | *SD* |
| Total sample | 8,400 | 12.90 | 1.96 | 8,600 | 85.72 | 3.66 | 8,150 | 16.69 | 1.51 | 700 | 47.32 | 1.97 | 8,250 | 17.51 | 2.30 |
| Child's race/ethnicity[2] | | | | | | | | | | | | | | | |
| White | 3,600 | 12.80 | 1.89 | 3,700 | 85.64 | 3.60 | 3,500 | 16.61 | 1.45 | 300 | 47.43 | 1.96 | 3,550 | 17.51 | 2.25 |
| Black | 1,300 | 12.86 | 2.00 | 1,350 | 85.75 | 3.68 | 1,300 | 16.86 | 1.56 | 200 | 47.20 | 1.99 | 1,300 | 17.45 | 2.40 |
| Hispanic, race specified | 1,150 | 13.09 | 2.04 | 1,200 | 85.99 | 3.79 | 1,100 | 16.75 | 1.59 | 100 | 47.37 | 1.97 | 1,150 | 17.64 | 2.26 |
| Hispanic, no race specified | 500 | 13.34 | 2.08 | 500 | 85.87 | 3.77 | 500 | 16.95 | 1.50 | 50 | 46.58 | 2.18 | 500 | 18.06 | 2.58 |
| Asian | 850 | 12.45 | 1.86 | 850 | 85.14 | 3.59 | 800 | 16.50 | 1.53 | # | 47.18 | 0.85 | 850 | 17.16 | 2.10 |
| Native Hawaiian/ Pacific Islander | 50 | 12.99 | 1.83 | 50 | 84.93 | 4.93 | 50 | 17.05 | 1.71 | # | 47.72 | 0.63 | 50 | 18.11 | 1.72 |
| American Indian/ Alaska Native | 200 | 13.14 | 1.96 | 250 | 85.45 | 3.87 | 250 | 17.17 | 1.65 | 0 | † | † | 200 | 17.92 | 2.01 |
| More than 1 race | 650 | 12.66 | 1.84 | 650 | 85.53 | 3.40 | 600 | 16.42 | 1.48 | 50 | 47.71 | 1.58 | 650 | 17.23 | 2.07 |
| Poverty status | | | | | | | | | | | | | | | |
| Below poverty threshold | 1,850 | 13.06 | 2.08 | 1,900 | 85.38 | 3.72 | 1,800 | 16.82 | 1.53 | 150 | 46.94 | 1.88 | 1,850 | 17.88 | 2.54 |
| At or above poverty threshold | 6,500 | 12.85 | 1.92 | 6,700 | 85.81 | 3.64 | 6,350 | 16.65 | 1.50 | 500 | 47.44 | 1.98 | 6,400 | 17.41 | 2.22 |

See notes at end of table.

Table 7-1. Children's average physical measurements and standard deviations for total sample and by key demographic variables, 2-year data collection: 2003–04—Continued

| | Children's average physical measurements | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Weight | | | Height | | | Middle upper arm circumference | | | Head circumference[1] | | | Body mass index | | |
| Characteristic | Number | Mean (kg) | *SD* (kg) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) | Number | Mean | *SD* |
| Child's sex | | | | | | | | | | | | | | | |
| Male | 4,300 | 13.22 | 1.93 | 4,400 | 86.18 | 3.70 | 4,150 | 16.81 | 1.51 | 350 | 47.71 | 1.93 | 4,200 | 17.76 | 2.32 |
| Female | 4,100 | 12.56 | 1.93 | 4,200 | 85.23 | 3.56 | 4,000 | 16.56 | 1.49 | 350 | 46.93 | 1.92 | 4,050 | 17.25 | 2.26 |
| Child's age at assessment | | | | | | | | | | | | | | | |
| 21 months and under | # | 11.39 | 1.39 | # | 84.65 | 3.70 | # | 17.8 | 0.60 | # | ‡ | ‡ | # | 15.91 | 1.82 |
| 22–23 months | 800 | 12.70 | 2.02 | 850 | 84.69 | 3.39 | 800 | 16.69 | 1.57 | 50 | 47.28 | 1.58 | 800 | 17.64 | 2.28 |
| 24–25 months | 6,450 | 12.86 | 1.94 | 6,650 | 85.57 | 3.57 | 6,300 | 16.66 | 1.50 | 500 | 47.32 | 1.99 | 6,350 | 17.52 | 2.31 |
| 26–27 months | 850 | 13.18 | 1.90 | 900 | 87.27 | 3.60 | 850 | 16.84 | 1.47 | 100 | 47.37 | 1.96 | 850 | 17.27 | 2.15 |
| 28 months and over | 200 | 13.89 | 2.23 | 250 | 88.84 | 4.42 | 200 | 17.07 | 1.54 | # | 47.34 | 2.30 | 200 | 17.59 | 2.66 |
| Birth weight | | | | | | | | | | | | | | | |
| Normal | 6,200 | 12.97 | 1.95 | 6,350 | 85.89 | 3.61 | 6,000 | 16.72 | 1.51 | † | † | † | 6,100 | 17.54 | 2.31 |
| Moderately low | 1,300 | 12.08 | 1.82 | 1,350 | 83.88 | 3.48 | 1,250 | 16.38 | 1.43 | † | † | † | 1,300 | 17.14 | 2.15 |
| Very low birth | 850 | 11.42 | 1.91 | 900 | 82.03 | 4.05 | 850 | 16.15 | 1.59 | 700 | 47.32 | 1.97 | 850 | 16.89 | 2.24 |
| Mother's age (in years) | | | | | | | | | | | | | | | |
| 19 and under | 300 | 13.15 | 2.13 | 300 | 85.87 | 3.64 | 250 | 16.80 | 1.58 | 50 | 47.03 | 1.64 | 300 | 17.81 | 2.58 |
| 20–29 | 3,750 | 12.94 | 1.99 | 3,850 | 85.63 | 3.61 | 3,650 | 16.76 | 1.51 | 300 | 47.14 | 1.87 | 3,650 | 17.61 | 2.37 |
| 30–39 | 3,750 | 12.82 | 1.87 | 3,800 | 85.80 | 3.68 | 3,550 | 16.60 | 1.48 | 250 | 47.53 | 2.07 | 3,600 | 17.39 | 2.17 |
| 40 and over | 650 | 12.93 | 2.19 | 650 | 85.75 | 3.84 | 650 | 16.68 | 1.57 | 50 | 47.44 | 2.01 | 650 | 17.43 | 2.44 |

See notes at end of table.

Table 7-1. Children's average physical measurements and standard deviations for total sample and by key demographic variables, 2-year data collection: 2003–04—Continued

| | Children's average physical measurements | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Weight | | | Height | | | Middle upper arm circumference | | | Head circumference[1] | | | Body mass index | | |
| Characteristic | Number | Mean (kg) | *SD* (kg) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) | Number | Mean | *SD* |
| Mother's race/ethnicity[2] | | | | | | | | | | | | | | | |
| White | 3,950 | 12.80 | 1.88 | 4,100 | 85.64 | 3.60 | 3,850 | 16.61 | 1.45 | 350 | 47.44 | 1.93 | 3,900 | 17.42 | 2.24 |
| Black | 1,350 | 12.86 | 2.00 | 1,400 | 85.78 | 3.68 | 1,350 | 16.84 | 1.56 | 200 | 47.24 | 2.07 | 1,300 | 17.44 | 2.38 |
| Hispanic, race specified | 1,400 | 13.26 | 2.10 | 1,450 | 86.04 | 3.80 | 1,350 | 16.83 | 1.59 | 100 | 47.09 | 1.98 | 1,400 | 17.86 | 2.42 |
| Hispanic, no race specified | 50 | 13.21 | 1.87 | 50 | 85.51 | 3.29 | 50 | 17.08 | 1.73 | # | ‡ | ‡ | 50 | 18.03 | 1.98 |
| Asian | 1,000 | 12.39 | 1.85 | 1,050 | 85.12 | 3.57 | 1,000 | 16.47 | 1.55 | # | 47.51 | 1.36 | 1,000 | 17.11 | 2.11 |
| Native Hawaiian/ Pacific Islander | 50 | 13.29 | 1.72 | 50 | 87.14 | 3.43 | 50 | 17.38 | 1.61 | # | ‡ | ‡ | 50 | 17.49 | 1.70 |
| American Indian/ Alaska Native | 300 | 12.96 | 2.03 | 300 | 85.23 | 3.64 | 300 | 16.91 | 1.58 | # | ‡ | ‡ | 300 | 17.80 | 2.18 |
| More than 1 race | 250 | 12.50 | 1.97 | 250 | 85.06 | 3.53 | 200 | 16.47 | 1.39 | # | 47.42 | 1.57 | 250 | 17.13 | 2.06 |

See notes at end of table.

Table 7-1. Children’s average physical measurements and standard deviations for total sample and by key demographic variables, 2-year data collection: 2003–04—Continued

| | Children’s average physical measurements | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Weight | | | Height | | | Middle upper arm circumference | | | Head circumference[1] | | | Body mass index | | |
| Characteristic | Number | Mean (kg) | *SD* (kg) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) | Number | Mean | *SD* |
| Mother’s education | | | | | | | | | | | | | | | |
| 8th grade or below | 350 | 13.14 | 2.17 | 400 | 85.82 | 3.96 | 350 | 16.92 | 1.61 | 50 | 47.23 | 1.35 | 350 | 17.84 | 2.52 |
| 9–12th grades | 1,700 | 13.04 | 2.08 | 1,750 | 85.52 | 3.75 | 1,650 | 16.86 | 1.61 | 150 | 47.00 | 2.20 | 1,650 | 17.78 | 2.56 |
| High school diploma | 1,800 | 12.92 | 1.98 | 1,850 | 85.67 | 3.66 | 1,750 | 16.77 | 1.55 | 150 | 47.21 | 1.77 | 1,750 | 17.55 | 2.24 |
| Vocational/ technical | 150 | 12.82 | 2.03 | 200 | 85.84 | 3.37 | 150 | 16.81 | 1.52 | # | 47.89 | 1.65 | 150 | 17.38 | 2.06 |
| Some college | 2,050 | 12.86 | 1.91 | 2,100 | 85.79 | 3.60 | 2,000 | 16.62 | 1.43 | 150 | 47.48 | 1.99 | 2,000 | 17.44 | 2.27 |
| Bachelor’s degree | 1,350 | 12.73 | 1.86 | 1,400 | 85.83 | 3.60 | 1,350 | 16.44 | 1.40 | 100 | 47.78 | 1.86 | 1,350 | 17.24 | 2.05 |
| Graduate school (no degree) | 150 | 12.70 | 1.83 | 150 | 85.62 | 4.03 | 150 | 16.23 | 1.36 | # | 47.79 | 2.28 | 150 | 17.26 | 2.75 |
| Master’s degree | 600 | 12.72 | 1.67 | 600 | 85.74 | 3.47 | 550 | 16.58 | 1.37 | 50 | 47.14 | 1.64 | 550 | 17.26 | 1.99 |
| Doctoral/ professional degree | 200 | 12.83 | 1.69 | 200 | 86.32 | 3.26 | 200 | 16.62 | 1.41 | # | 46.92 | 2.12 | 200 | 17.19 | 1.80 |

† Not applicable.

# Rounds to zero.

‡Reporting standards not met; too few cases for analysis.

[1] Obtained from those born at very low birth weight only (1,500 grams or less).

[2] Race categories exclude Hispanic origin unless specified.

NOTE: The child weight W2C0 was applied to obtain these statistics, however the cell counts are unweighted to demonstrate the distribution in the ECLS-B at 2 years. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

Each trainer’s physical measurements were obtained in advance and standards for reliability were set. For example, if a trainer’s standard weight was determined to be 65 kg, any measurement between 61.75 kg and 68.25 kg would be considered reliable and, therefore, acceptable. Any value outside that range would signal that a third measurement was needed. Training staff collected each interviewer’s physical measurement forms at the end of the session and reviewed the measurements on each form. If the measurements in a set differed from the standard measurements by more than 5 percent and if a required third measure was not obtained, the interviewer was required to attend a help lab and to demonstrate competence to the trainer in obtaining the physical measurements. In addition, because a trainer or training assistant served as the focal child, the interviewer’s measurement procedures could be observed at the time and any errors in procedure addressed directly at the end of that trainee’s turn. The purpose of the certification, therefore, was to identify those interviewers having problems to make sure they were retrained before leaving training. In this way, by the end of training, all interviewers were certified on the physical measurements.

### 7.3.2 Reliability and Data Quality Control During the Data Collection Year

Quality control was also maintained on an ongoing basis during the year of data collection. As physical measurements data were entered at Westat, any meaningful (e.g., discrepancies greater than 5 percent and a required third measurement not obtained) or out-of-range errors were noted, and the interviewer’s field supervisor was notified immediately. The field supervisor then followed up with the interviewer and provided corrective feedback. This process was made possible by a fast feedback loop, ensuring that interviewers received timely feedback about any systematic errors.

As an indicator of reliability, correlations between the two measurements in a set were obtained. Table 7-2 presents the correlations between the first and second measurements within each set, as well as the means and standard deviations for these physical measurements. Because the point of this table is to summarize the reliability of the interviewers and not population estimates, these are unweighted statistics. There were some cases for which only a single measurement within a set was obtained, presumably due to lack of cooperation from the child. This was true for 1.3 percent (n = 113) of the child weight measurements, 1.4 percent (n = 122) of the child height measurements, 2.3 percent (n = 193) of the child MUAC measurements, and 2.8 percent (n = 30) of the child head circumferences measurements. In addition, reserve codes (i.e., -9, -8, and -7) have been deleted from these analyses. For further information about how the physical measurements composites were created and how differences larger than 5 percent were treated, please refer to chapter 7 of the *User's Manual for the ECLS-B Longitudinal 9-Month–2-Year Data File and Electronic Codebook* (NCES 2006-046) (Nord et al. 2006).

Table 7-2. Reliability of sets of physical measurements, 2-year data collection: 2003–04

| Variable | Mean | Standard deviation | Correlation (*r*) |
|---|---:|---:|---:|
| Child weight | | | |
| C2CHWGT1 (n=8,650) | 12.61 kg | 1.97 kg | .99* |
| C2CHWGT2 (n=8,550) | 12.61 kg | 1.96 kg | (n=8,550) |
| Child height (n=8,800) | | | |
| C2CHHGT1 (n=8,900) | 85.12 cm | 3.83 cm | .99* |
| C2CHHGT2 (n=8,800) | 85.16 cm | 3.83 cm | (n=8,800) |
| Middle upper arm circumference | | | |
| C2MUAC1 (n=8,450) | 16.58 cm | 1.54 cm | .99* |
| C2MUAC2 (n=8,250) | 16.58 cm | 1.53 cm | (n=8,250) |
| Head circumference | | | |
| C2CHHC1 (n=1,100) | 47.81 cm | 1.98 cm | .98* |
| C2CHHC2 (n=1,050) | 47.80 cm | 1.99 cm | (n=1,050) |

* $p < .05$.
NOTE: Sample sizes have been rounded to the nearest 50.
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.

## 7.4 Comparison of 2-Year and 9-Month Physical Measurements

To compare the 2-year physical measurements with the 9-month measurements, the difference between the composites was obtained by subtracting the 9-month composite from the 2-year composite for each measure, for example, X2CHHGT–X1CHLENG to yield overall increase in height, and applying the round 2 longitudinal child weight W2C0. To obtain these differences, all cases with missing data were omitted and the difference had to be greater than or equal to zero (i.e., cases with a round 2 value smaller than the round 1 value were deleted from the analysis). (For further information about cases that had 2-year physical measurements that were smaller than those obtained at 9-months, please see the Data Anomalies section of the *User's Manual for the ECLS-B Longitudinal 9-Month–2-Year Data File and Electronic Codebook* [NCES 2006-046] [Nord et al. 2006].) These increases in growth are presented in the following table, grouped by key demographic variables. Because BMI could not be obtained at 9 months, it is not included in this table.

Table 7-3. Average growth as measured by increases in physical measurements, 2-year and 9-month data collections: 2001–02 and 2003–04

| | Average growth in physical measurements from 9-month to 2-year data collection | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Weight | | | Height | | | MUAC | | | Head circumference[1] | | |
| Characteristic | Number | Mean (kg) | *SD* (kg) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) |
| Total sample | 8,000 | 3.49 | 1.68 | 8,550 | 12.68 | 3.69 | 5,550 | 1.69 | 1.78 | 600 | 3.50 | 2.15 |
| Child's race/ethnicity[2] | | | | | | | | | | | | |
| White | 3,450 | 3.45 | 1.64 | 3,700 | 12.66 | 3.65 | 2,350 | 1.65 | 1.76 | 250 | 3.26 | 1.93 |
| Black | 1,250 | 3.41 | 1.61 | 1,350 | 12.74 | 3.79 | 900 | 1.71 | 1.79 | 150 | 3.65 | 2.00 |
| Hispanic, race specified | 1,100 | 3.53 | 1.77 | 1,200 | 12.81 | 3.72 | 750 | 1.79 | 1.82 | 100 | 3.73 | 1.97 |
| Hispanic, no race specified | 500 | 3.81 | 1.87 | 500 | 12.71 | 3.93 | 3,200 | 1.93 | 1.88 | 50 | 4.72 | 3.68 |
| Asian | 800 | 3.36 | 1.58 | 850 | 12.34 | 3.51 | 550 | 1.66 | 1.65 | # | 3.11 | 1.19 |
| Native Hawaiian/ Pacific Islander | 50 | 3.37 | 1.24 | 50 | 11.75 | 3.07 | 50 | 1.39 | 1.90 | # | 3.59 | 0.40 |
| American Indian/ Alaska Native | 200 | 3.03 | 1.54 | 250 | 11.04 | 4.17 | 150 | 1.74 | 1.77 | 0 | † | † |
| More than 1 race | 600 | 3.44 | 1.56 | 650 | 12.67 | 3.39 | 50 | 1.45 | 1.75 | 50 | 3.00 | 2.58 |
| Poverty status | | | | | | | | | | | | |
| Below poverty threshold | 1,800 | 3.61 | 1.90 | 1,850 | 12.39 | 3.96 | 1,250 | 1.84 | 1.85 | 150 | 3.64 | 2.00 |
| At or above poverty threshold | 6,200 | 3.45 | 1.61 | 6,650 | 12.76 | 3.61 | 4,300 | 1.65 | 1.76 | 450 | 3.46 | 2.19 |
| Child's sex | | | | | | | | | | | | |
| Male | 4,100 | 3.48 | 1.68 | 4,350 | 12.36 | 3.69 | 2,800 | 1.64 | 1.79 | 300 | 3.57 | 2.42 |
| Female | 3,900 | 3.49 | 1.68 | 4,200 | 13.02 | 3.67 | 2,750 | 1.74 | 1.77 | 300 | 3.44 | 1.84 |
| Child's age at assessment | | | | | | | | | | | | |
| 21 months and under | # | 2.45 | 1.18 | # | 15.85 | 3.90 | # | 1.48 | 0.73 | # | ‡ | ‡ |
| 22–23 months | 800 | 3.45 | 1.64 | 850 | 12.58 | 3.30 | 550 | 1.59 | 1.70 | 50 | 3.56 | 2.03 |
| 24–25 months | 6,150 | 3.47 | 1.67 | 6,600 | 12.62 | 3.64 | 4,300 | 1.67 | 1.77 | 450 | 3.39 | 2.02 |
| 26–27 months | 800 | 3.57 | 1.68 | 850 | 13.15 | 4.11 | 550 | 1.98 | 1.98 | 50 | 3.97 | 2.82 |
| 28 months and over | 200 | 3.98 | 2.09 | 200 | 13.17 | 4.94 | 150 | 1.79 | 1.77 | # | 4.94 | 2.23 |
| Birth weight | | | | | | | | | | | | |
| Normal | 5,900 | 3.50 | 1.68 | 6,300 | 12.66 | 3.69 | 4,100 | 1.69 | 1.78 | † | † | † |
| Moderately low | 1,250 | 3.37 | 1.59 | 1,300 | 12.87 | 3.75 | 850 | 1.74 | 1.80 | † | † | † |
| Very low | 800 | 3.33 | 1.60 | 850 | 13.00 | 3.93 | 600 | 1.54 | 1.66 | 600 | 3.50 | 2.15 |

See notes at end of table.

Table 7-3. Average growth as measured by increases in physical measurements, 2-year and 9-month data collections: 2001–02 and 2003–04—Continued

| | Average growth in physical measurements from 9-month to 2-year data collection | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Weight | | | Height | | | MUAC | | | Head circumference[1] | | |
| Characteristic | Number | Mean (kg) | *SD* (kg) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) | Number | Mean (cm) | *SD* (cm) |
| Mother's age (in years) | | | | | | | | | | | | |
| 19 and under | 250 | 3.77 | 1.97 | 300 | 12.86 | 3.83 | 200 | 1.61 | 1.64 | 50 | 2.93 | 1.65 |
| 20–29 | 3,550 | 3.50 | 1.73 | 3,800 | 12.52 | 3.75 | 2,500 | 1.71 | 1.80 | 300 | 3.61 | 2.38 |
| 30–39 | 3,700 | 3.44 | 1.57 | 3,950 | 12.80 | 3.62 | 2,500 | 1.66 | 1.75 | 250 | 3.47 | 1.96 |
| 40 and over | 450 | 3.61 | 1.84 | 500 | 12.87 | 3.69 | 300 | 1.85 | 1.97 | 50 | 3.32 | 1.68 |
| Mother's race/ethnicity[2] | | | | | | | | | | | | |
| White | 3,750 | 3.45 | 1.62 | 4,050 | 12.66 | 3.60 | 2,600 | 1.64 | 1.75 | 300 | 3.34 | 2.07 |
| Black | 1,250 | 3.43 | 1.62 | 1,350 | 12.80 | 3.79 | 950 | 1.74 | 1.83 | 200 | 3.49 | 1.85 |
| Hispanic, race specified | 1,350 | 3.66 | 1.86 | 1,400 | 12.76 | 3.87 | 900 | 1.78 | 1.82 | 100 | 3.96 | 2.70 |
| Hispanic, no race specified | 50 | 2.80 | 1.81 | 50 | 13.13 | 3.20 | 50 | 3.02 | 1.98 | # | ‡ | ‡ |
| Asian | 950 | 3.29 | 1.54 | 1,000 | 12.35 | 3.48 | 650 | 1.69 | 1.72 | # | 3.01 | 1.20 |
| Native Hawaiian/ Pacific Islander | 50 | 3.23 | 1.21 | 50 | 12.31 | 3.48 | 50 | 1.61 | 1.74 | # | ‡ | ‡ |
| American Indian/ Alaska Native | 300 | 3.09 | 1.51 | 300 | 11.18 | 3.98 | 200 | 1.42 | 1.60 | # | ‡ | ‡ |
| More than 1 race | 200 | 3.47 | 1.54 | 250 | 12.42 | 3.71 | 150 | 1.50 | 1.80 | # | 3.58 | 2.51 |
| Mother's education | | | | | | | | | | | | |
| 8th grade or below | 350 | 3.58 | 1.84 | 350 | 12.50 | 3.66 | 250 | 1.72 | 1.72 | # | 4.90 | 2.91 |
| 9–12th grades | 1,600 | 3.58 | 1.94 | 1,700 | 12.46 | 4.19 | 1,200 | 1.88 | 1.94 | 150 | 3.24 | 2.19 |
| High school diploma | 1,650 | 3.50 | 1.69 | 1,800 | 12.56 | 3.60 | 1,200 | 1.68 | 1.75 | 150 | 3.55 | 2.04 |
| Voc./technical | 150 | 3.60 | 1.61 | 150 | 12.36 | 3.04 | 100 | 1.32 | 1.10 | # | 3.81 | 1.18 |
| Some college | 1,950 | 3.47 | 1.54 | 2,050 | 12.67 | 3.61 | 1,350 | 1.55 | 1.71 | 150 | 3.55 | 2.15 |
| Bachelor's degree | 1,300 | 3.40 | 1.53 | 1,400 | 13.03 | 3.42 | 900 | 1.75 | 1.80 | 100 | 3.65 | 2.08 |
| Graduate school (no degree) | 150 | 3.32 | 1.32 | 150 | 13.59 | 2.96 | 100 | 1.08 | 1.02 | # | 3.11 | 0.71 |
| Master's degree | 550 | 3.34 | 1.44 | 600 | 12.85 | 3.47 | 350 | 1.64 | 1.74 | 50 | 3.03 | 2.31 |
| Doctoral/professional degree | 200 | 3.36 | 1.59 | 200 | 13.21 | 3.56 | 100 | 1.74 | 1.84 | # | 3.00 | 1.24 |

† Not applicable.

# Rounds to zero.

‡Reporting standards not met; too few cases for analysis.

[1] Obtained from those born at very low birth weight only (1,500 grams or less).

[2] Race categories exclude Hispanic origin unless specified.

NOTE: The round 2 weight W2C0 was used to obtain these statistics; however the cell counts are unweighted to demonstrate the distribution in the ECLS-B at 2 years. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

*This page is intentionally left lank.*

# 8. TODDLER'S SECURITY OF ATTACHMENT STATUS

The development of attachment relationships between children and parents is one of the most important aspects of socioemotional development in the infancy to toddler period. Although most major theories of socioemotional development have contributed to our understanding of parent-child relationships, attachment theory has become the predominant paradigm over the past 2 decades and has generated a large volume of research. This chapter provides a brief overview of attachment theory, a discussion of possible attachment measures, and the decision to develop a brief attachment measure for the Early Childhood Longitudinal Survey, Birth Cohort (ECLS-B). This is followed by a description of the Toddler Attachment Sort-45 Item (TAS-45) and its place in the direct child assessment. The TAS-45 obtains rich data, and the variables obtained are discussed and summarized at the end of this chapter.

## 8.1 Overview of Attachment Theory

Attachment is the deep emotional bond that forms between a young child and the parents and forms the basis for the child's development of a sense of security. When a young child feels secure, he or she is more likely to explore the environment freely, returning to the parent for comfort when distressed. Researchers have shown (Waters et al. 1995) that the effects of early attachment are enduring, even into adulthood.

Children's formation of secure attachments with caregivers is a hallmark of socioemotional growth and development in the infancy to toddler period (Lamb 2000; Main 2000). According to attachment theorists (Bowlby 1969, 1973, 1980), the attachment relationship has its roots in the earliest months of life, in what is called the *indiscriminate social responsiveness phase.* This theory claims that babies are born with a repertoire of *pre-programmed* signals that elicit care from and closeness with caregivers, usually the mother or father, or both. These attachment behaviors are those that promote the child's safety and survival and include crying, grasping, clinging, smiling, vocalizing (e.g., cooing), reaching, and crawling. In return, the parents are highly motivated to engage in reciprocal attachment behaviors that are geared to relieve the causes of distress, such as picking up the child, cuddling the child, providing comfort, and soothing the child and to perpetuate the positive aspects of the interaction, such as smiling, eye-gazing, talking, etc.

Through repeated interactions, the parents' sensitivity and responsiveness help the baby develop more mature and better organized neural control mechanisms, and therefore gain greater self-regulatory capacities. Associated with this is the emergence of what has been called *stranger anxiety*, which marks the child's emerging ability to recognize specific individuals and differentiate them from the parents. A preference for interacting with the parents over strangers is the hallmark of the second phase, called *discriminating sociability*, which typically occurs by the age of 7 months. During this second phase, the young child's emotional bond with the parent(s) increases and deepens as perceived needs are met by an emotionally available caregiver. An important characteristic of this phase is that the young child develops a sense of trust in the parent, who can be depended on to respond appropriately to the child's signals.

The third phase of the development of an attachment relationship extends from about 7 months through 2 years and is the focus of the TAS-45. The hallmark of this phase is the emergence of a specific attachment style and formation of a secure base. Attachment style is an important development because it becomes internalized into a working model of the self and of the self in relationship with others. That internalized working model, in turn, is carried forward into adult life and can become core belief systems about the world.[1] The formation of secure attachments in early life is somewhat like a protective factor in that the child is able to use the parent as a secure base from which to explore novel stimuli in the environment freely, acquire a sense of self-confidence and adaptability to new and challenging situations, and to focus sustained attention without interruption from anxiety or anger. According to Grossmann (Grossman, Grossman, and Zimmerman 1999), "secure attachment provides the best-known psychological precondition for tension-free playful exploration."

Insecure attachment arises when the parents (or other primary caregivers) are not emotionally available to the child, lack affection for the child, or are unable to interpret the needs of the child, and, therefore, are unable to comfort the distressed child. There are generally two types of insecure attachment that have been recognized by researchers over the past 2 decades: the *ambivalent attachment style* (also known as anxious/resistant) and the *avoidant attachment style* (Lamb 2000; Main 2000).

Ambivalent attachment results when parents are inconsistent and unpredictable with the child, responding sensitively sometimes and not at other times. As a result, the child cannot trust the parent's availability. This often shows up at extreme distress when separated from the parent and inability

[1] For example, a child who is responded to sensitively and effectively will experience him/herself as worthy of care and will experience the world as a benevolent place, whereas a child whose needs are not consistently met (or whose needs are thwarted) will conclude that he/she is not worthy of care and that life is unhappy and uncomfortable.

to be comforted by the parent upon reunion. Such children also manifest anxiety about moving away from the parent to explore the environment, leading to a lack of independence.

Avoidant attachment arises when parents are outwardly rejecting and respond to the child with hostility or indifference. Children with this style are at risk of growing up to avoid others and deny their own needs. Such children may look independent to others but in actuality they appear independent because they have not learned to depend on others (Lamb 2000; Main 2000).

In the past 2 decades, researchers have found evidence for a disorganized/disoriented style of attachment that seems to result when parents are outwardly abusive toward the child or simply neglectful. According to researchers, children with this style seem to be a blend of both the avoidant and ambivalent attachment styles. Hypersensitive to perceived abuse, such children may appear to become confused or disoriented in the presence of the parent, trying one moment to please the parent and the next moment displaying anger to or rejection of the parent (Main 2000).

In sum, researchers have identified four common styles of attachment:

- Attachment Style A: Avoidant;
- Attachment Style B: Secure;
- Attachment Style C: Ambivalent (also called anxious/resistant); and,
- Attachment Style D: Disorganized/disoriented.

The attachment literature is fairly clear about the distribution of the three main types of attachment styles, secure (the “B’s”), avoidant (the “A’s”) and ambivalent (the “C’s”). Researchers working with separate data sets and in different cultures have generally found that most children (about 85 percent) can be classified into one of these three main styles. Of those, in general, approximately 60 percent of children can be categorized as B, securely attached, approximately 20 to 25 percent can be categorized as A, avoidant, and approximately 10 to 15 percent can be categorized as C, ambivalent in their attachment (Ainsworth et al. 1978). Furthermore, the shape of this distribution is similar across cultures, although specific percentages of A’s and C’s may vary (van IJzendoorn and Sagi 1999, p. 729). In addition, in the United States, approximately 15 percent of children are difficult to classify into one of these three styles. Main and Solomon (1986), have characterized these children as having a “disorganized/disoriented” attachment style, characterized by a lack of a coherent attachment strategy for interacting with the parent.

Children’s security of attachment in the first 2 years has been systematically related to variations in maternal caregiving behavior and to children’s subsequent outcome measures. Secure attachment is related to higher cognitive and social functioning, higher levels of self-esteem, better peer skills, and greater *ego-resilience* (i.e., a personality strength that enables an individual to develop an adaptive self that responds flexibly to novel situations; Block and Block 1980) during toddlerhood. Moreover, attachment classifications of children are consistently correlated with maternal responsiveness, competence, and maternal self-confidence. Attachment security has also been linked to more positive marital adjustment and other qualities of the marital relationship, and to levels of social support provided to the parent by family members and friends (Hazan and Zeffman 1999; Feeney 1999). Thus, the nature and degree of the child’s attachment seems to be important not only as precursor to the child’s later socioemotional development but also as an indicator of the quality of the parent-child relationship that contributes to individual differences in growth trajectories.

Including a measure of children’s attachment in the ECLS-B 2-year data collection fits well with the use of the Nursing Child Assessment Teaching Scale (NCATS) at 9 months and the Two Bags Task at 2 years, both of which measure characteristics of parent-child interaction (e.g., parental sensitivity, children’s engagement cues, and positive regard for the parent) that are associated with attachment formation and other outcomes, such as developmental status, vocabulary growth, and self-regulation skills. Because the ECLS-B is the first national study to follow children from infancy to school age, the addition of an in-depth measure of attachment greatly enhances the validity of the data and the usefulness of the findings. Many research studies, including large-scale national studies, such as the Early Head Start national evaluation, the Comprehensive Childcare Development Project, and the National Institute of Child Health and Human Development (NICHD) Early Child Care Study have included measures of attachment spanning the age range of 12 months through 3 years.

The two most common assessments of attachment appropriate for use with children at about 2 years of age are the Strange Situation (Ainsworth et al. 1978), a complex laboratory-based situation not suitable for the ECLS-B, and the Attachment Q-Sort (AQS) (Waters and Deane 1985), which was developed as an alternative to the Strange Situation. The TAS-45 which was adapted from the AQS, describes children’s attachment security, dependency, and sociability on the basis of observations made in the home. The advantages of the AQS include observations done in naturalistic settings and its applicability in a range of countries, such as China, Colombia, Germany, Israel, Japan, Canada, Norway, and the United States (Posada et al. 1995).

However, the AQS has 90 items and uses a Q-sort procedure, a methodology that has been used in psychological research for several decades (Block 1961) but which is new to survey research. A Q-sort can be used to measure a wide range of characteristics, such as personality traits, consumer preferences, and preschooler behavior problems. In a traditional Q-sort, items are printed on a card (the size of a business card), one item per card. The individual completing the sort then reviews each card and places it into one of several piles. One pile might be for items characteristic of a construct, another pile for items that are not characteristic of that construct, and a third pile of items that are neutral (neither characteristic nor uncharacteristic). The individual then does a second layer of sorting to further define the items into those that are highly characteristic, characteristic, somewhat characteristic, neutral, somewhat uncharacteristic, uncharacteristic, and highly uncharacteristic.

Researchers usually place limitations on the final placement of items into piles so that, for example, the distribution of items in piles should resemble a normal distribution or should be a flat distribution with equal numbers of items in all piles. This adds a layer of complexity to the sorting procedure and requires additional time to work through the details of taking items out of fuller piles and finding places for them in smaller piles, without compromising the description of the object of measurement.

Although it was deemed important to include a measure of attachment in the 2-year data collection of the ECLS-B, neither the Strange Situation nor the AQS was appropriate for application in a large-scale field setting. A third study, the Study of Early Child Care, had included a set of 40 questions about children’s separation and reunion behaviors when they are dropped off and picked up from day care. The measurement of children’s separation and reunion behaviors is frequently considered a proxy measure of children’s status of attachment. However, at the time of the design of the 2-year data collection, the results of this set of items were not yet publicly available. This set of 40 separation and reunion questions was tested for use in the ECLS-B during early rounds of cognitive testing during the design phase of the parent interview. This cognitive testing showed that this set of questions took respondents about 10 minutes to complete during a parent interview that was already too long and, therefore, was not feasible for the ECLS-B at 2 years.

Dr. Brian Vaughn of Auburn University, a Technical Review Panel (TRP) member, recommended contacting a noted authority in attachment research, Dr. John Kirkland of Massey University. Dr. Kirkland was asked to explore the possibility of developing a shortened AQS for the 2-year ECLS-B data collection. Using AQS datasets acquired from researchers in many different countries, Dr. Kirkland and his colleague Dr. David Bimler had been using Multidimensional Scaling and facet cluster analysis to identify the AQS items that obtained the best information possible about children's security of attachment.

Dr. Kirkland's work to develop a shortened version of the AQS, the TAS-45, is described in the next section, followed by a brief description of how the TAS-45 was administered in the context of the home visit at 2 years. The items included in the TAS-45 and the related variables on the data file are then summarized, as well as the training procedures used to train field staff and the procedures used to maintain reliability during the data collection year. A key indicator for the validity of the TAS-45 would be variability with demographic variables and with such outcome measures as children's Bayley Short Form–Research Edition (BSF-R) mental and motor scale scores. For this reason, the final section includes a table of TAS-45 classifications for the total sample and for the major demographic grouping variables.

## 8.2 Development of the TAS-45 Items and Adaptation to a Laptop Application

The AQS (Waters and Deane 1985) is a 100-item measure that is Q-sorted into 9 piles. A revised version published in 1995 (Waters 1995) has 90 items, only 45 of which were redundant with the 1985 version, for a total pool of 145 unique items. Each item is a description of children's behaviors with the mother under stressful circumstances, such as when a friendly stranger is in the room. The AQS uses the Q-sort procedure that was described above. In a research context, the AQS requires at least 3 hours of in-home observation of a child before completing the sort. Sorting 100 items into 9 piles can take up to another 45 minutes to complete.

There were several constraints on the administration of an attachment assessment in the ECLS-B (i.e., limited time available for the home visit, the need to reduce interviewer and respondent burden, and the need to include simple measures that do not require extensive training of interviewers) that made use of the AQS unfeasible. In contrast, interviewers in the ECLS-B needed to complete a measure of attachment in less than 10 minutes on the basis of observations made during the 90-minute home visit. Therefore, the shortened and procedurally streamlined version was developed, the TAS-45, which had 45 items that were sorted into four piles, with roughly equal numbers of items sorted into each pile. To further streamline the procedure for the 2-year ECLS-B, a laptop application of the sorting procedure was developed. This laptop application was completed by the interviewer after the home visit was completed. Because the home visit was in the range of 2 or more hours, the interviewer had ample opportunity to observe the child behaviors covered in the TAS-45.

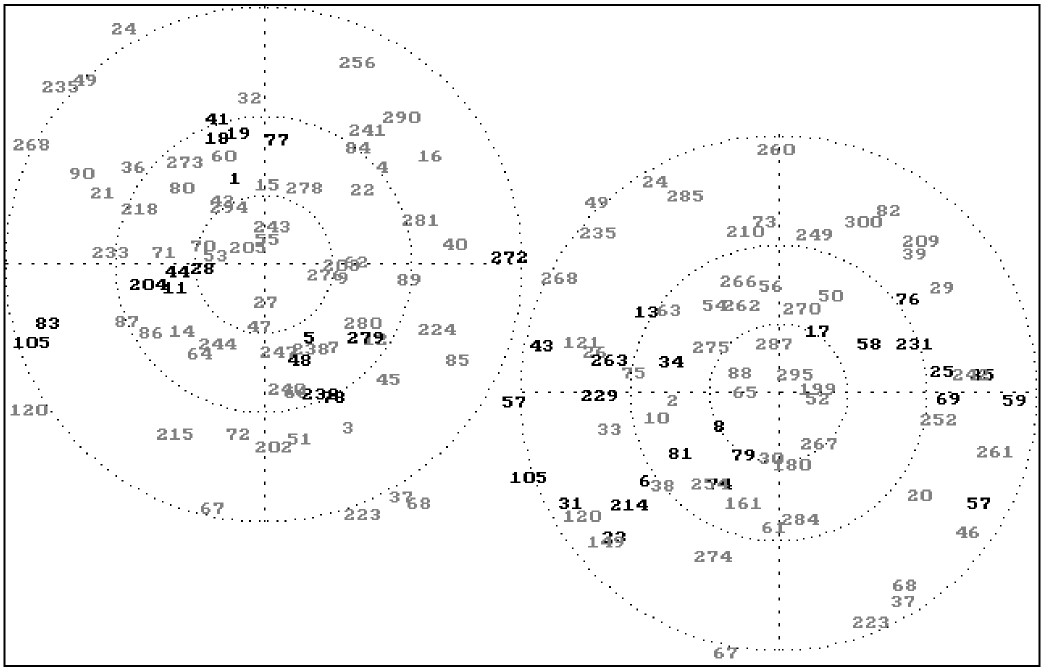
The work to develop the TAS-45 proceeded on four fronts: identifying the best subset of AQS items to include in a shortened version; shortening the Q-sorting procedure by reducing the number of piles in the final level of placements; fashioning a laptop application to reduce burden to respondents and interviewers; and identifying an additional subset of six items that would best capture the disorganized style of attachment. Based on research, it was expected that 10-15 percent of children in the United States would fit in the disorganized style of attachment category. These disorganized style children will require a greater proportion of social services due to such issues as failure to thrive (Ward, Lee, and Lipper 2000), externalizing behavior (e.g., conduct disorder) problems (Greenberg, DeKlyen, Speltz, and Endriga 1997; Speltz, Greenberg, and DeKlyen 1990) and internalizing (e.g., depression) problems (Moss, St. Laurent, and Parent 1999). Therefore, it is important for the first national study of child development to obtain information that allows for the identification of children with disorganized attachment styles, in keeping with the study’s goal of obtaining comprehensive information about children’s experiences and characteristics.

### 8.2.1 Identifying the Core Attachment Items

Attachment researchers in the United States, Colombia, Japan, Germany, Israel, Canada, and Norway provided Dr. Kirkland with several hundred subjective Q-sort datasets. Dr. Kirkland then used multidimensional scaling (MDS), followed by facet cluster analysis on these datasets to map all the items from the AQS (i.e., the 100-item version from 1985 and the 90-item version from 1995 [minus the redundant items]). To oversimplify, imagine a map of the United States with one AQS item residing in New York City, another in Seattle, another in Boston, etc. MDS measures the distances between the items and locates them on a map. Then further imagine a number of items centered around New York City at varying distances from each other, and so on. Facet cluster analysis then identified the points of congregation (or centers, e.g., New York City) where items characteristic of a dimension are clustered. As a result of the MDS analyses the locations of the items and their points of congregation (or centers) formed a three dimensional rendering outcome from the MDS space and was roughly spherical.

The map itself provides a frame upon which these subjective data are spread and interpreted. Psychologically meaningful nodes are revealed in the map as points labeled *hotspots*, described by surrounding items. As a result, instead of X possible dimensions (where X = number of items) specifying a person's location on the map, it becomes possible to display a particular person's subjective sort-data as a profile, by calculating weights for each salient hotspot. Typically there are fewer than 10 useful hotspots. Further, once hotspots have been established, it is possible to create alternate forms of any instrument, providing an equivalent number of items are available in the immediate vicinity of each hotspot. With specific regard to the ECLS-B, this meant that if an item was not feasible for the field setting, it could be easily replaced with another item from the same hotspot. In exhibit 8-1,[2] each item is represented by its original AQS item number.

Exhibit 8-1. Map of the 145 unique Attachment Q-Sort items



With the redundant items eliminated, Dr. Kirkland mapped the remaining 145 items into eight clusters, or dimensions, which could be said to describe attachment behaviors: comfortably cuddly, cooperative, enjoys company, independent, attention-seeking, upset by separation, avoids others/not sociable, and demanding. From these 145 items, a subset of 39 of the items with the strongest associations to the eight dimensions, with approximately four to six items per dimension, were selected. (The 39

[2] This exhibit was prepared by the developer of the TAS-45, Dr. John Kirkland of Massey University, Palmerston North, New Zealand, in 1999.

selected items are highlighted in boldface in exhibit 8-1.) As work progressed through field testing, it was found that some of these items needed to be replaced by other items from the same hotspot due to difficulties with wording, observability in the field or difficulties for interviewers.

Representative items from the 39 selected items include the following (a complete list of TAS-45 items can be found in appendix B):

- When mother asks child to do something, child understands what she wants (may or may not obey).
- When child cries, cries loud AND long.
- A social child who enjoys the company of others.
- Turns away from friendly adult strangers (i.e., the interviewer) if they come too close.
- If asked, lets friendly adult strangers (i.e., the interviewer) hold or share toys.

Abbreviating the number of items in the sort was a first step toward making the TAS-45 feasible for administration in the field. Another step was to shorten the procedure and streamline it so that interviewers untrained in attachment theory and research would be able to complete the task reliably and in a reasonable amount of time.

### 8.2.2 Shortening the Q-Sort Procedure

To complete the full AQS, the sorter begins with a stack of 100 or 90 cards depending on the version used, one item printed per card. The sorter reads through the cards on a first pass and sorts them into three piles; those that apply, those that do not apply, and those that are in the middle, meaning the sorter is not yet sure or has insufficient evidence to place it anywhere else. The sorter then re-sorts each pile twice more so that on the last step there are 9 piles of items that range from “highly characteristic” of the child to “highly uncharacteristic” of the child. The items are either distributed evenly across piles or according to an algorithm based on the normal curve (e.g., 18 cards in the middle pile), tapering down to 5 cards in each outermost pile.

This type of sorting procedure would not have been feasible for the ECLS-B because it would take too long for respondents to sort and count, sort and recount, etc. An alternative method advocated by Dr. Kirkland was the Method of Successive Sorts (MOSS) (Block 1961), in which the sorter begins by sorting the items into two piles, the *applies pile,* and the *not-applies pile.* Each pile is then re-sorted in turn to produce a total of eight piles ranging form *applies most* to *applies least.* The MOSS procedure also has the advantage that it is not necessary to monitor how many cards are in each pile. However, even the MOSS procedure was too time-consuming for the ECLS-B. To remedy this problem, Dr. Kirkland conducted analyses that determined that sorting to the second level (e.g., to a total of four piles ranging from “almost always applies” to “rarely or hardly ever applies” to the child [with a fifth pile as a holding place for the *undecideds*]), was sufficient to obtain reliable data. Correlations between the four- and nine-pile solutions for each item ranged from 0.95 to 0.99. Therefore, it was decided to use the two-step, four-pile sorting procedure, because it was feasible for the field setting of the ECLS-B.

### 8.2.3 The 39-Item TAS in the 18-Month Field Test

The 4-pile, 39-item version of the TAS (TAS-39) was tested during the May 2001 18-month field test. Two TAS-39 sorts were obtained. One was completed by the interviewer immediately after the home visit. The other sort was completed by the parent while the interviewer administered the BSF-R to the child. The interviewer instructed the parent about how to complete the sort before beginning the BSF-R.

The parent-completed sort was obtained in 74 percent of cases and the interviewer-completed sort in 98 percent of cases. The rate of 25 percent noncompletion by parents was higher than expected and was probably due to multiple issues. For one, parents who were unable to read the items and follow the instructions were unable to complete the TAS-39. For parents with poor reading skills, it was difficult to understand the subtleties of some items. Other parents had difficulty following the written instructions on the sorting sheet (a laminated sheet that provided an illustration of the sorting procedure accompanied by written instructional boxes). Finally, some parents did not complete the sort because they were so engrossed watching the child complete the BSF-R that they neglected to complete the TAS-39. On the basis of the field test, it was decided to drop the parent-completed sort.

The following variables were derived for each child separately for the parent-completed sort and for the interviewer-completed sort: the eight hotspots and the traditional A-B-C classifications. On the basis of attachment research using the Strange Situation (typically used from 12-18 months of age), the attachment literature is fairly clear about the distribution of the three main types of attachment styles, secure (the B’s), avoidant (the A’s) and ambivalent (the C’s). Researchers working with separate datasets and in different cultures have in general found that approximately 60 to 70 percent of very young children can be categorized as securely attached (the B's), approximately 20 to 25 percent can be categorized as avoidant (the A's), and approximately 10 to 15 percent can be categorized as ambivalent in their attachment (the C's).

The distribution of children across attachment categories that resulted from the interviewer-completed sorts was generally in line with research findings in the attachment literature. For this reason it was concluded that interviewers could observe and sort children's attachment behaviors successfully, as summarized in table 8-1. However, the distribution shows fewer securely attached children and a slightly higher frequency of avoidant children.[3] This is probably due to simple differences between the samples of subjects typically recruited in an academic research setting (i.e., upper middle class, White parents, with high levels of education) with the nationally representative sample of the ECLS-B field test.

Table 8-1. Percentage distribution of children's attachment classifications obtained by interviewer-completed TAS-39 sorts in the 18-month field test: 2001

| Attachment classification | Expected distribution[1] | Interviewer-completed sort |
|---|---:|---:|
| A (avoidant) | 20 | 28 |
| B (secure) | 70 | 58 |
| C (ambivalent) | 10 | 14 |

[1] Expected distribution is based on "Strange Situation" procedures typically done at 12 to 18 months and scored by trained observers. n=617
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 18-month field test, 2001.

### 8.2.4 Identifying the Subset of Disorganized Items

Simultaneous with the 2001 field test, Dr. Kirkland acquired additional datasets from researchers investigating the disorganized style of attachment and subjected these data to a meta-analysis in order to map the items characteristic of the disorganized/disoriented attachment style. While estimates of the proportion of young children who exhibit the disorganized/disoriented style of attachment behavior suggest that the percentage of children in this category is far from negligible (by one account ranging from 14 percent in middle-class, nonclinical North American groups to 24 percent in low socioeconomic samples; Lyons-Ruth and Jacobvitz 1999), an accurate estimate based on a large scale, nationally representative sample has never been obtained.

[3] A chi-square test comparing the expected and observed distributions of children across the attachment classifications indicates that the observed and expected distributions are significantly different from one another ($\chi^2 = 42.15$, degrees of freedom = 2, $p < .05$.)

Children with this attachment style show a disorganized response to the parent, especially when distressed. For example, a child may freeze in fear when approaching the parent as if torn between wanting contact but fearful of the parent's response. Researchers believe that this disorganized style is highly associated with negative outcomes. Disorganized children, who some researchers estimate to be from 10 to 15 percent of children, place heavy demands on special education, community, and mental health services. Because this population is of particular concern with regard to policy planning and implementation, it was important that the ECLS-B capture information on this attachment classification. Dr. Kirkland's analyses determined that an additional six items would be sufficient to identify disorganized children who were not easily classified as avoidant, secure, or ambivalent.

From an initial pool of 42 items characterizing the disorganized attachment style, Dr. Kirkland, again using multidimensional scaling and facet cluster analysis of aggregated datasets, narrowed the selection to 12 items that were strongly associated with the disorganized style. Many of the D items would have been difficult for field staff to observe, either because of their subtlety or because of the distress they may cause to the observer (e.g., disoriented and finds it hard to focus when near mother, or, stiffens up when held by mother). Such observations could be distressing if an interviewer recognizes the implications of such behaviors for the child's well-being. Working with Westat staff, these 12 items were reviewed with an eye toward objectivity and feasibility in the field, and a final set of 6 disorganized items was identified and added to the original TAS-39.

Representative items that characterize the disorganized style include the following:

- Comes to mother to give her toys but will not touch or look at her.
- Goes all floppy (limp) when held by mother.
- Suddenly aggressive toward mother for no reason (e.g., hits, slaps, pushes, bites mother).
- Generally cranky or grouchy when with mother.
- With mother, child suddenly switches mood. For instance, goes from being nice to mean, or calm to upset (crying, afraid, angry), or gets upset and then goes blank.
- Looks dazed and unsure (e.g., stares blankly, or freezes in an unusual position for a few seconds).

In sum, a 45-item TAS was designed that can be completed with just 4 piles (plus 1 for unsure) that obtains reliable data in about 10 minutes by field interviewers on the basis of approximately 90 minutes of observation. Once these items were identified, Westat child development staff worked with Dr. Kirkland to make sure that the items were observable in the context of the home visit, that the difficulty of the language used for the items was at about the sixth to eighth grade level, and that field staff could be trained to observe the target behaviors.

## 8.3 TAS-45 Protocol in the 2-Year Data Collection

The TAS-45 is an observational measure that was completed by the interviewer after the completion of the home visit. From the point of view of the respondent during the home visit, the TAS-45 was virtually invisible. It was completed entirely by the interviewer on the basis of observations made during the home visit, particularly during the direct child assessments. The interviewer completed the sort on the laptop application that was built into the Child Observation section of computer-assisted personal interview (CAPI), which only became available to the interviewer if the direct child assessments had been completed. In the Child Observations, the interviewer first completed the Interviewer Observations about the child's behavior during the BSF-R and then completed the observations about the child's home environment. Upon completion of these two sets of items, the interviewer was prompted to complete the TAS-45. The results of each sort were stored on the interviewer's laptop in an ASCII data file and transmitted back to the home office during routine transmittal calls. There was no need, therefore, to have the TAS-45 data entered by computer-assisted data entry staff at the home office.

## 8.4 TAS-45 Variables: Hotspots and Traditional Classifications

The advantage of the TAS-45 is that it generates a rich set of data. Researchers who simply want to be able to classify children according to their predominant style of attachment can use the A-B-C and D classification, X2TASCLS. 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.

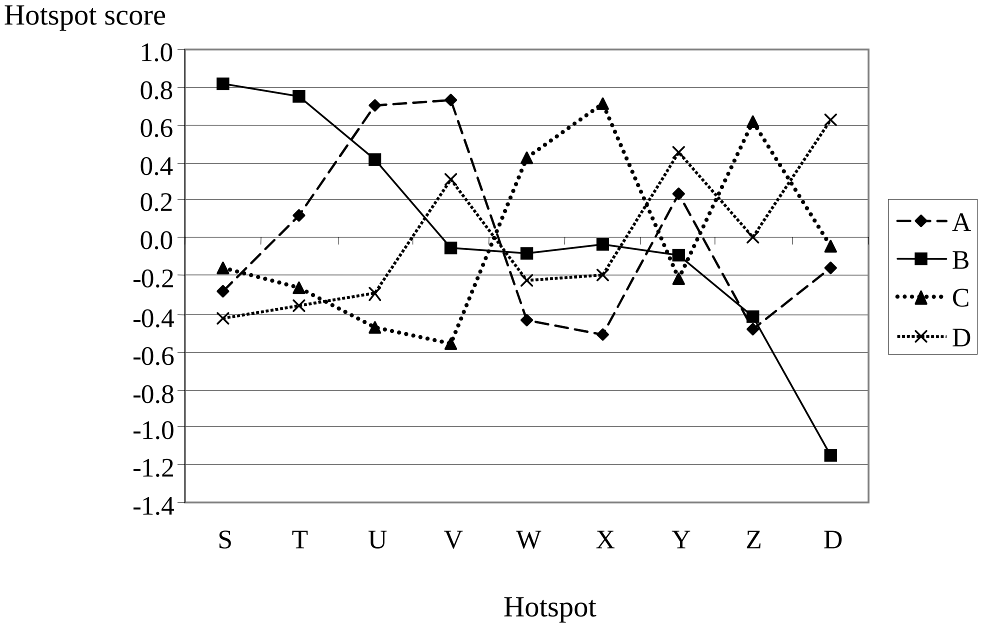
Each item in the TAS-45 has a hotspot weight that represents the contribution of the item to the total score for that hotspot. The closer an item is to the center of a cluster, the greater its weight. Conversely, the more distant an item is from the center of a cluster, the smaller the weight. These weights form a continuous scale from some maximum value for items immediately adjacent to that hotspot, down to 0 for irrelevant ones. The weights are easily calculated in a computational spreadsheet. Briefly, this spreadsheet consists of a column for each hotspot, containing constants derived from the earlier meta-analysis of the aggregated datasets obtained from several different attachment researchers. The item placements from a particular sort is followed by multiplying each item's entry in the spreadsheet by a value for that row, which is simply the number of the pile to which the corresponding item was assigned (ranging from applies most to applies least). The products are totaled down the columns, producing a set of nine hotspot scores (i.e., the weighted sum of the items in the hotspot), which can be plotted as a line-graph that illustrates an individual's profile of attachment behaviors. Minor refinements to this process take into account missing data and items not used in the sort.

On the basis of an individual's hotspot scores, a profile can be plotted that characterizes the individual's attachment style. The developer of the TAS-45 identified the three types of hotspot profiles that are associated with the dominant classification types. Children's profiles on the hotspots were then classified into the four types by using the correlation of the child's profile with the ideal profiles and assigning the classification with the highest correlation (provided that the correlation was greater than IrI = 0.40). In those cases where there was no correlation greater than $r = 0.40$, Euclidean distance between the child's profile and the ideal profiles was used with assignment of the classification determined by the shortest distance. The confidence level, X2TASCNF, described in more detail below, is an indicator of how confident the analyst can be that the assigned classification is the best description of the child's attachment style.

Each of the four different styles has a characteristic profile. The four profiles are shown in figure 8-1[4]: one is classified as attachment style B, secure; one as A, avoidant; one as C, ambivalent; and one as D, disorganized. This highlights the importance of the profile of the hotspot scores, in determining the classification.

[4] This figure was prepared by the developer of the TAS-45, Dr. John Kirkland of Massey University, Palmerston North, New Zealand, in 2005.

Figure 8-1. Characteristic profiles of typical attachment styles: Attachment Type B, secure (solid line), Attachment Type A, avoidant (dashed lines), Attachment Type C, ambivalent (dotted line), and Attachment Type D (small-dotted line)



Key: S = Warm and cuddly
T = Cooperative
U = Enjoys company (sociable)
V = Independence

W = Attention seeking
X = Upset by separation
Y = Avoids others
Z = Demanding, angry
D = Moody, unsure, unusual

In figure 8-1, the darker solid lines illustrate Attachment Type B, secure, which is high on warmth, cooperativeness, and sociability, and low on separation distress and avoidance of others. Attachment Type A, avoidant, is illustrated by the dashed lines, which are high on sociability and independence but low on warmth and cooperativeness. Attachment Type C, ambivalent, is low on sociability and independence and high on attention-seeking and separation distress. Attachment Type D, disorganized is high on the disorganized behaviors and somewhat high on avoidance and independence and low on the more sociable behavior. The hotspots and their descriptions are summarized in table 8-2.

The TAS-45 is extremely rich with respect to the data obtained. Consistent with the traditional attachment classifications, the TAS-45 is able to generate the classical A (avoidant), B (secure), and C (ambivalent) categories consistent with the categories obtained from the original Strange Situation, to which the D (disorganized) category is added.

Table 8-2. Variable names and descriptions for the TAS-45 hotspots, traditional classification types, TAS classification confidence score, and traditional AQS security and dependency scores obtained in the 2-year ECLS-B data collection: 2003–04

| Variable name | TAS-45 name | Description of construct |
|---|---|---|
| Hotspots | | |
| X2TASHS1 | TAS Hotspot 1:<br>Warm, cuddly | Child actively seeks and enjoys physical affection with the parent, whether or not child is distressed. The score ranges from -1 to 1, with a 1 meaning that the child engages in warm and cuddly behavior quite often and -1 meaning that the child rarely, if at all, engages in such behaviors. |
| X2TASHS2 | TAS Hotspot 2:<br>Cooperative | Child is compliant and cooperative with parental requests and suggestions. The score ranges from -1 to 1, with a 1 meaning that the child often displays cooperative behavior in interaction with the parent and a -1 meaning that the child rarely, if at all, displays such cooperation. |
| X2TASHS3 | TAS Hotspot 3:<br>Enjoys company | Child is sociable and enjoys the company of others. The score ranges from -1 to 1, with a -1 meaning that the child rarely, if at all expresses enjoyment when in the company of others and a 1 meaning that the child often approaches others and enjoys interacting with others. |
| X2TASHS4 | TAS Hotspot 4:<br>Independence | Child is independent and self-sufficient, explores freely. The score ranges from -1 to 1, with 1 meaning that the child is often independent and self-sufficient and -1 meaning that the child rarely, if ever, engages in independent activity. |
| X2TASHS5 | TAS Hotspot 5:<br>Attention seeking | Child needs to be center of parent’s attention; child demands attention. The score ranges from-1 to 1 with 1 meaning that the child often demands the parent’s attention and -1 meaning that the child rarely demands attention. |
| X2TASHS6 | TAS Hotspot 6:<br>Upset by separation | Child becomes upset when mother is out of sight; child is inconsolable without mother. The score ranges from -1 to 1 with 1 meaning that the child is very easily upset by any separation from the mother and -1 meaning that the child does not become upset when the mother moves out of sight or leaves the room |
| X2TASHS7 | TAS Hotspot 7:<br>Avoids others | Child prefers inanimate objects; avoids people; is slow to warm up to strangers. The score ranges from -1 to 1, with 1 meaning that the child frequently avoids other individuals and prefers to focus on inanimate objects (e.g., toys) and -1 meaning that the child rarely, if ever, engages in such avoidant behaviors. |

See note at end of table.

Table 8-2. Variable names and descriptions for the TAS-45 hotspots, traditional classification types, TAS classification confidence score, and traditional AQS security and dependency scores obtained in the 2-year ECLS-B data collection: 2003–04—Continued

| Variable name | TAS-45 name | Description of construct |
|---|---|---|
| Hotspots—Continued | | |
| X2TASHS8 | TAS Hotspot 8: Demanding, angry | Child is quick to become angry to get own way, e.g., if parent is unresponsive; is quick to cry; is slow to stop crying. The score ranges from -1 to 1, with 1 meaning that the child becomes angry and demanding if the parent does not respond to the child's requests immediately and -1 meaning that the child rarely, if ever, becomes angry when the parent does not respond immediately. |
| X2TASHS9 | TAS Hotspot 9: moody, unsure, unusual | Child displays unusual behaviors, has quick mood changes; looks confused or dazed. The score ranges from -2 to 2 on this variable. A score of 2 means that the child has demonstrated one or more unusual behaviors, such as hitting the mother for no apparent reason, going limp when held by the mother, or rapidly changing from one mood (e.g., calm) to another (e.g., rage) for no apparent reason. A score of -2 means that the child did not demonstrate any of these behaviors. |
| Traditional Classification Scores | | |
| X2TASCLS | TAS Classification: A, B, C, or D | Classic security of attachment categories, consistent with "Strange Situation" measure.<br>Attachment Type A, avoidant<br>Attachment Type B, secure<br>Attachment Type C, ambivalent (sometimes called anxious/resistant)<br>Attachment Type D, disorganized |
| X2TASCNF | TAS Confidence in classification | Measure of confidence in X2TASCLS, it measures the distance between the individual's attachment profile and the closest prototypical A-B-C profile. The shorter the distance, the more confidence in the classification (however, the higher the value of X2TASCON, the more confidence we can have in the classification. |

See note at end of table.

Table 8-2. Variable names and descriptions for the TAS-45 hotspots, traditional classification types, TAS classification confidence score, and traditional AQS security and dependency scores obtained in the 2-year ECLS-B data collection: 2003–04—Continued

| Variable name | TAS-45 name | Description of construct |
|---|---|---|
| Traditional Classification Scores—Continued | | |
| X2TASSEC | TAS Security Factor Score | This is the traditional Security factor score obtained by the AQS and is obtained in the same method by using published criterion sorts for the Security construct (Waters and Deane 1985). The combination of Security scores and Dependency scores in relation to each other also points to the traditional classification type. A low (or minus) security score and high dependency score suggest Attachment Type C, ambivalent; A low security (or minus) score plus a low (or minus) dependency score suggest Attachment Type A, avoidant; and a high (positive) security score and low (negative) dependency score suggest Attachment Type B, secure. |
| SECurity and DEPendency Scores | | |
| X2TASDEP | TAS Dependency Factor Score | This is the traditional Dependency factor score obtained by the AQS and is obtained in the same method by using published criterion sorts for the Dependency construct (Waters and Deane 1985). |

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.

There is also a classification confidence (X2TASCNF) score that indicates to the analyst how well each child's classification summarizes the hotspot scores. X2TASCNF ranges from 0 to 1. A high value (closer to 1) indicates that the assigned attachment classification is the best descriptor for that individual's hotspot profile. A low value (closer to 0) indicates that a particular child's hotspot profile is less well captured by the classification, or that the profile blends two classifications, one only slightly more than the other. A confidence value in the range of 0.30 or lower would be an indication to review the individual child's pattern of hotspot scores to examine whether the classification is the best description or whether the hotspot profile suggests an alternative classification. The decision based on the confidence level is not whether or not to eliminate cases with low confidence scores, but whether to use the classification score X2TASCLS, or the hotspot scores, or use the hotspot scores in addition to the classification. For example, there is a small subset of cases in which X2TASCNF is in the region of 0.30 to 0.35 and in which the assigned classification is Attachment Type D, disorganized. The confidence level is low for these cases because they seem to be equidistant between the profiles for Attachment Type B, secure, and Attachment Type D, disorganized.

In those cases where the value of the confidence score is low, another alternative is to investigate using the Security (X2TASSEC) and Dependency (X2TASDEP) scores instead of the TAS attachment classification. These scores are traditionally obtained by researchers using the AQS and are factor scores. 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, with scores closer to -1 indicating low ability to use the adult as a secure base (the security score), or low clinginess (the dependency score). These scores can each be used on their own to examine concurrent associations, for example between security and exploratory competence, or dependency and fearfulness. Because -1 is a valid score for these two variables, the conventional use of reserve codes to indicate missing data does not apply to these variables, so that only -9 is used to indicate missing data for these variables.

In addition, the TAS-45 generates scores for the nine hotspots (or clusters) described in table 8-3, which are characteristics used to create the profiles of children's attachment styles. The hotspots provide more detailed information about children's attachment behaviors than just the classifications or the security and dependency scores. The traditional classifications were derived from observations of children's behaviors during the Strange Situation (Ainsworth, Blehar, Waters and Wall, 1978), which is a laboratory-based situation designed to elicit attachment behaviors from children aged 12 to 18 months. The hotspots obtain information that captures the greater repertoire of children's attachment behavior at 2 years, as well as information on more age-appropriate behaviors.

While the hotspot scores are used to create profiles of attachment style, it is recommended that analysts use the traditional attachment classifications when examining issues related to attachment. 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, such as "Enjoys company," and measures of social functioning in subsequent data collections. As another example, a researcher could examine the associations between X2TASH52, cooperative, and subsequent outcome measures of adjustment to school. The hotspot values on the data file are linear and ordinal. Strictly speaking, they are proportionate in nature and not of the "classic" Likert-type. The range of values for the hotspot variables is from -1 to 1, although the range for hotspot 9 is between -1 and 2 because the developer of the TAS-45 wanted this hotspot to be more sensitive than the others because of its diagnostic importance for indicating possible problems in the attachment system of the parent and child.

## 8.5 TAS-45 Training, Certification, and Quality Control

It was important that that all trainees received consistent training to ensure the reliable interpretation of children’s behaviors and that trainees continued to be reliable during the year of data collection to prevent *coder drift* or a change over time in how information is collected. For example, due to fatigue, a shift in perspective that accrues from experience, or simply forgetting of the coding instructions, the individual relaxes the coding standards so that reliability of coding is compromised. The TAS-45 items needed to be presented as objectively and consistently as possible to make sure that interviewers all recognized the same behaviors. To address both of these challenges a computer-based training (the TAS CBT) was designed by Dr. Kirkland and his colleague Andrew Drawneek, for the 2-year national training.

A key consideration of the design of the TAS CBT was that trainees NOT be introduced to the theory behind attachment security. This decision was made for two reasons. First, field staff do not have extensive backgrounds in child development and it would be time-consuming to present attachment theory. Training time would be better spent training them to recognize the behaviors they would see during home visits than in understanding the underlying theory imperfectly, at best. Rather, field staff was told that the TAS-45 focused on the kinds of behaviors children use to obtain the kind of care they want and need. Second, it was necessary to minimize the potential for bias on the part of the field staff who may have become distressed when observing behaviors that they knew had negative implications for the child. Field staff may be less willing to recognize the behaviors identified in the TAS-45 items if they think they are indicative of future difficulties or pathology. Therefore, the TAS CBT focused on learning to recognize the TAS-45 items, to understand the subtle differentiations between items, and to describe the children’s behaviors using the items.

As mentioned earlier, the TAS CBT was designed to train interviewers on the TAS-45, thereby providing them with the skills to complete the sorting procedure quickly and efficiently. The TAS CBT was loaded onto enough laptops so that each trainee could complete the TAS CBT prior to arriving at training. The laptops with the TAS CBT were sent to trainees several weeks before the scheduled training so that they could complete the TAS CBT prior to arrival at training. The TAS CBT was accompanied by a 20-page hard-copy manual that presented a general overview of concepts and the items included in the sort, described the sorting procedure, provided hard-copy examples of the sorting procedure, and explained the modular approach of the TAS CBT.

The TAS CBT had three modules, which the trainees had to complete in order. At the end of each module, the trainee completed a brief quiz. The quiz had to be passed at a minimum of 80 percent in order for the trainee to advance to the next module. If the trainee did not pass the quiz, that module needed to be redone until a passing score was achieved. The 80 percent minimum for passing was considered adequate by Dr. Kirkland. Based on his experiences conducting attachment research, 80 percent agreement at the item level would assure adequate reliability for the measure as a whole.

The goal of the first module was to familiarize trainees with the 45 items and to familiarize them with the click-and-drag technique for moving items into the placement piles. The module concluded with an exercise in which the trainee sorted the items into their proper categories.

Module 2 had two goals. One goal was to familiarize trainees with the more subtle differences between items. Trainees' success on this module was assessed with an odd-one-out exercise in which three items at a time were presented on the screen, and the trainee had to indicate which one of the three was least like the others.

The second goal of this module was to familiarize trainees with the sorting procedure. Three brief written vignettes (provided also in hard copy) were presented on the laptop, which the trainee read and then completed a sort for by clicking and dragging each item, placing it into the pile that best described the child in the vignette. For example, the trainee could place an item, such as "Cries loud AND long," into the pile characterized as "most like" the child.

The third module included videoclips of three children, each of whom typified one of the three main styles of attachment: secure, avoidant, and anxious/resistant. In order to sensitize trainees to the secure base behaviors that children use to maintain proximity with the mother, and to the behaviors that enable children to leave their mothers to explore the environment, the first videoclip was silent. This served to highlight the behavioral aspects of children's attachment styles without the distraction of the dyad's conversation. The remaining videoclips had full audio.

Because this methodology was new, the construct required good observational skills that rely on familiarity with the items, and the sorting procedure required practice, a 2-hour block of time was reserved at the national training for trainees to repeat this entire three-module training package. Attendance at this session was mandatory. During this mandatory session, trainees were encouraged to ask questions and discuss any difficulties they were having. Therefore, trainees had a minimum of 5 hours of training on the TAS-45, 2 hours at training, and an average of at least 3 hours at home.

Upon completion of the national training, results of the module quizzes were downloaded from each trainee's laptop. These results were in the form of a flat ASCII file that was then sent to the developer of the TAS-45, who then calculated agreement rates by comparing the trainees' results with standardized results obtained from graduate and undergraduate students who were studying attachment theory and who were known to be reliable on the TAS-45. The key measure of reliability was the agreement rate for the final videoclip in module three. The average agreement rate for the ECLS-B field staff was 82 percent, which exceeded the 80 percent minimum.

After national training, videotapes of a simulated child assessment portion of the home visit were sent to field staff. They were instructed to watch the videotape and then complete a sort based on the child behaviors they observed. The profiles from these sorts were then compared with profiles that were obtained by the developer of the TAS-45.

The videotapes that interviewers watched were made by three members of Westat's Child Development staff who made simulated home visits to several volunteer mother-child pairs. During these home visits, one staff member administered the entire direct assessment protocol while the other staff member videotaped it. Upon return to the home office, these two individuals reviewed the tape and did a preliminary TAS-45 sort to try to identify one child from each of the major attachment classification types: Attachment Type A, avoidant; Attachment Type B, secure; and Attachment Type C, ambivalent. It was deemed important to assess interviewers' reliability across the classifications, because it was hypothesized that some attachment behaviors (e.g., the behaviors associated with the secure classification) may be more salient than others, for example, the avoidant style, making it easier to complete a sort for those classifications.

Copies of each of the three selected videotapes were then sent to Dr. Kirkland at Massey University. He and his students (who were studying attachment theory) then completed sorts to classify the attachment types of the children in the videotapes and developed prototype profiles for each of the children. These prototype profiles were then used to evaluate the reliability of the ECLS-B interviewers based on how well their sort profiles for each child resembled these prototypical profiles.

The first videotape, sent out 3 months after the national training, was of a child classified by Child Development staff as an Attachment Type B (securely attached). Three months later, the second videotape, of a child classified as an Attachment Type C (ambivalent), was sent to interviewers. The third videotape, of a child classified as an Attachment Type A (avoidant), was sent out approximately 8 months into the data collection. It was intended that all four attachment types be included in the reliabilities. However, it proved impossible to identify and recruit a child who could be classified as an Attachment Type D (disorganized). Therefore, reliability on Type D could not be obtained.

After each reliability videotape was sent out to the Westat interviewers, they were instructed to complete the reliability quality control form for the BSF-R first and then rewind the tape and view it again, this time paying attention to the child's attachment domain behaviors. They then completed the TAS-45 and returned the results to the Westat home office.

The result of each completed sort was then transmitted in an ASCII data file to Dr. Kirkland to obtain reliabilities for each interviewer. Results of these reliability checks showed that some attachment styles are easier, and, therefore, more reliable, to sort than others. As was expected, Attachment Type B (secure) was the most reliable, with interviewers averaging 88 percent agreement with the prototypical profiles. Because the majority of children can be classified as secure, it is particularly important that they be sorted reliably. On the second videotape, that showed a child classified as Attachment Type C (ambivalent), interviewers averaged 83 percent agreement with the prototypical profiles. On the third videotape, that showed a child classified as Attachment Type A (avoidant), the agreement rate was only 75 percent, which is a bit low. Avoidant behaviors were difficult for the interviewers to recognize because many avoidant behaviors also resemble independence. Overall, however, interviewers averaged 82 percent agreement on the three reliability videotapes, which the developer of the TAS-45 considered acceptable.

## 8.6 TAS-45 Results in the 2-Year National Data Collection

How well the TAS-45 performed in the 2-year national data collection can be determined by longitudinal associations between TAS-45 scores and key outcome measures, such as the BSF-R and, ultimately, school readiness. In addition, longitudinal associations between attachment precursors, such as NCATS scale scores, and 9-month BSF-R scores, with the TAS-45 at 2 years and with the 2-year BSF-R will identify developmental inequalities associated with emotional functioning and well-being.

Simple descriptives for the hotspot scores are presented in table 8-3 and the frequency distributions for the four types of attachment (A-B-C-D) are presented in table 8-4. In addition, associations between the TAS-45 and the BSF-R and Two Bags Task are summarized in appendix A.

Table 8-3. Weighted means and standard deviations for TAS-45 hotspot scores in the 2-year ECLS-B data collection: 2003–04

| Variable name | Hotspot description | Mean | Standard deviation |
|---|---|---:|---:|
| X2TASHS1 | Warm and cuddly | 0.30 | 0.30 |
| X2TASHS2 | Cooperative | 0.39 | 0.34 |
| X2TASHS3 | Enjoys company | 0.21 | 0.39 |
| X2TASHS4 | Independent | 0.16 | 0.31 |
| X2TASHS5 | Attention seeker | -0.08 | 0.24 |
| X2TASHS6 | Upset by separation | -0.14 | 0.24 |
| X2TASHS7 | Avoids others, does not socialize | -0.01 | 0.27 |
| X2TASHS8 | Demanding, angry | -0.13 | 0.25 |
| X2TASHS9 | Moody, unsure, unusual (disorganized) | -0.59 | 0.52 |

NOTE: The sample size for the hotspot scores is 8,750 cases (rounded to the nearest 50). All data weighted using the child weight,W2C0.
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.

Table 8-4. Weighted percentage distribution for the four types of attachment (variable X2TASCLS [traditional classification scores]) in the 2-year data collection of the ECLS-B: 2003–04

| Classification type | Percent |
|---|---:|
| Attachment Type A, Avoidant | 16.27 |
| Attachment Type B, Secure | 61.12 |
| Attachment Type C, Ambivalent | 8.91 |
| Attachment Type D, Disorganized | 13.46 |

NOTE: The sample size for the percentage distribution is 8,750 cases (rounded to the nearest 50). All data weighted using the child weight,W2C0.
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.

As mentioned in section 8.1, a general finding in the attachment literature is that approximately 55 to 65 percent of children can be classified as Attachment Type B (secure), which is consistent with the results of the ECLS-B. Another approximately 15 to 20 percent are classified as Attachment Type A (avoidant), again consistent with the results of the ECLS-B. Generally, between 10 and 15 percent are classified as Attachment Type C (ambivalent); in the ECLS-B the percentage of children classified as Type C fell outside this range, at 9 percent. To date there have been no large-scale studies of a nationally representative sample that have investigated the prevalence of children classified as Attachment Type D (disorganized). However, the research literature suggests that approximately 5 to 15 percent of children would be classified as having this type of attachment. The percentage identified as disorganized in the ECLS-B is consistent with this.

Table 8-5 presents descriptive information (weighted percentages) about the TAS-45 attachment classifications for the total sample and by the key demographic grouping variables.

Table 8-5. Weighted percentages of attachment classifications for the total sample and by key demographic grouping variables in the 2-year data collection: 2003–04

| Variable | Traditional classification of attachment type | | | |
|---|---|---|---|---|
| | A (Avoidant) | B (Secure) | C (Ambivalent) | D (Disorganized) |
| Total sample | 16.27 | 61.12 | 8.91 | 13.46 |
| Mother's race/ethnicity[1] | | | | |
| White | 14.64 | 65.32 | 8.02 | 12.02 |
| Black | 23.58 | 52.87 | 7.99 | 15.57 |
| Hispanic, race specified | 16.22 | 57.13 | 11.81 | 14.83 |
| Hispanic, no race specified | 7.83 | 50.70 | 12.41 | 29.07 |
| Asian | 13.40 | 61.72 | 11.92 | 12.96 |
| Native Hawaiian/Pacific Islander | 22.48 | 63.20 | 5.51 | 8.81 |
| American Indian/Alaska Native/ | 22.25 | 49.14 | 8.35 | 17.25 |
| More than 1 race | 17.80 | 53.87 | 5.27 | 23.06 |
| Poverty status | | | | |
| Below poverty threshold | 20.12 | 51.58 | 10.85 | 17.45 |
| At or above poverty threshold | 15.27 | 63.92 | 8.40 | 12.41 |
| Child's race/ethnicity[1] | | | | |
| White | 14.31 | 65.70 | 8.08 | 11.91 |
| Black | 24.41 | 51.95 | 8.03 | 15.61 |
| Hispanic, race specified | 15.99 | 57.83 | 10.27 | 15.90 |
| Hispanic, no race specified | 16.66 | 56.88 | 12.84 | 13.62 |
| Asian | 13.64 | 61.70 | 12.22 | 12.44 |
| Native Hawaiian/Pacific Islander | 26.82 | 62.91 | 4.17 | 6.11 |
| American Indian/Alaska Native | 21.79 | 49.53 | 9.14 | 19.55 |
| More than 1 race | 17.61 | 58.99 | 6.65 | 16.74 |

See notes at end of table.

Table 8-5. Weighted percentages of attachment classifications for the total sample and by key demographic grouping variables in the 2-year data collection: 2003–04—Continued

| Variable | Traditional classification of attachment type | | | |
|---|---|---|---|---|
| | A (Avoidant) | B (Secure) | C (Ambivalent) | D (Disorganized) |
| Child's age at assessment | | | | |
| 21 months and under | 27.80 | 39.04 | 2.93 | 30.23 |
| 22–23 months | 13.40 | 63.82 | 8.92 | 13.86 |
| 24–25 months | 16.45 | 60.71 | 9.14 | 13.69 |
| 26–27 months | 18.53 | 61.84 | 7.75 | 11.88 |
| 28 months and over | 15.68 | 65.60 | 7.26 | 11.47 |
| Child's sex | | | | |
| Male | 17.88 | 54.94 | 9.71 | 17.47 |
| Female | 14.67 | 67.89 | 8.11 | 9.33 |
| Birth weight | | | | |
| Normal | 16.24 | 61.68 | 8.75 | 13.34 |
| Low | 17.72 | 56.79 | 10.55 | 14.95 |
| Very low | 14.48 | 53.69 | 14.18 | 17.66 |
| Mother's education | | | | |
| 8th grade and under | 16.03 | 63.18 | 10.28 | 10.52 |
| 9th–12th grades | 18.67 | 50.76 | 11.28 | 19.29 |
| High school diploma or equivalent | 17.48 | 59.57 | 8.85 | 14.10 |
| Vocational/technical program | 14.19 | 64.98 | 8.49 | 12.34 |
| Some college | 16.52 | 62.30 | 8.58 | 12.79 |
| Bachelor's degree | 14.63 | 66.98 | 7.32 | 11.07 |
| Graduate or professional school, no degree | 11.70 | 77.50 | 3.12 | 7.68 |
| Master's degree | 9.62 | 78.23 | 7.61 | 4.55 |
| Doctorate or professional degree | 15.60 | 61.92 | 7.70 | 14.78 |

See notes at end of table.

Table 8-5. Weighted percentages of attachment classifications for the total sample and by key demographic grouping variables in the 2-year data collection: 2003–04—Continued

| Variable | Traditional classification of attachment type | | | |
|---|---|---|---|---|
| | A (Avoidant) | B (Secure) | C (Ambivalent) | D (Disorganized) |
| Mother’s age (in years) | | | | |
| 19 and under | 16.68 | 48.63 | 11.38 | 23.31 |
| 20–29 | 18.06 | 57.79 | 9.24 | 14.91 |
| 30–39 | 14.62 | 65.53 | 8.48 | 11.37 |
| 40 and over | 14.57 | 64.68 | 8.68 | 12.06 |

[1] Race categories exclude Hispanic origin unless specified.

NOTE: The child weight W2C0 was applied to produce these statistics.

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

# 9. INTERVIEWER OBSERVATIONS OF THE CHILD AND THE CHILD'S ENVIRONMENT

At the 2-year data collection there were three parts to the interviewer observations: (1) a set of items about the child's behavior during the Bayley Short Form–Research Edition (BSF-R) assessment; (2) a set of interviewer-completed items about the child's home environment, supplemented by four questions from the parent interview; and (3) the Toddler Attachment Sort-45 item (TAS-45) discussed in chapter 8.[1] The first part was completed by the interviewer in the Child Observations portion of computer-assisted personal interviewing (CAPI) after the home visit. The second part, on the child's home environment, was completed by both the interviewer in the Child Observations portion of CAPI and by the parent respondent during the parent interview. The last part, the TAS-45 laptop Q-sorting application built into the Child Observations portion of CAPI, was completed by the interviewer after the home visit and after completing parts 1 and 2.

This chapter describes separately the first two parts of the Child Observations and how they compare with their 9-month counterparts. Training procedures to ensure that interviewers understood the observation items are described. Cross-sectional associations between the items describing children's behavior during the BSF-R and their BSF-R outcome scores are presented from the 2-year data collection, as well as descriptive summary statistics of children's home environments as a whole and grouped by key demographic variables. Finally, longitudinal comparisons between items at 9 months and 2 years are presented separately for each set of items.

## 9.1 Interviewer Observations of the Child's Behavior During the BSF-R

The first part of the interviewer observations consisted of a subset of 13 questions selected for the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) from the Behavior Rating Scale (BRS) of the Bayley Scales of Infant Development, Second Edition (BSID-II), 11 of which were completed by the interviewer (table 9-1) and two by the parent (table 9-2). In the 9-month data collection, there were nine items for the interviewer to complete, plus the two completed by the parent. At 2 years, four items were added because children's behavioral repertoire is larger at this age and because these additional behaviors become more predictive with development. Therefore, it is more likely that children will exhibit these additional behaviors at 2 years than at 9 months.

[1] The 9-month data collection incorporated only the first two parts for the interviewer observation.

The ECLS-B selected items that were representative of and developmentally appropriate for the target age, and that were easily rated by field interviewers. In other words, items were selected from a range of behaviors that interviewers were likely to observe in children of this age and in the context of the BSF-R. Items that were too clinical and, therefore, too difficult or too subjective to score were avoided. Items were not selected with the intention of creating a subscale of BRS items. Only a small subset of discrete behaviors from a range of domains was selected. Therefore, these interviewer observation items should not be considered the same as the BRS, nor should they be treated like a subscale of the BRS. To compare the BRS items in the 2-year ECLS-B with the full BRS used in the BSID-II, see the BSID-II manual (Bayley 1993). For further information about the 9-month BRS items, please refer to the *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) Methodology Report for the Nine-Month Data Collection (2001–02), Volume 1: Psychometric Characteristics* (Andreassen and Fletcher 2005).

After completing the home visit, including observing the child's behaviors during the BSF-R, the interviewer completed the BRS observational items in the Child Observations section of CAPI. The BRS items are rated on a 5-point scale based on the frequency of the observed behavior, sometimes combined with the qualitative aspect of the item, such as intensity or valence. For each item, the points on the scale are well described. For example, for the item "Adaptation to change in test materials," the scale ranges from "(1) Consistently resists relinquishing materials and/or refuses to accept new materials" at the low end to "(5) Consistently relinquishes materials and accepts new materials" at the high end.

The BRS items in the ECLS-B at 2 years also included the two questions asked of the parent respondent during the 9-month data collection. These two questions were included in the Child Activity Booklet. At the end of the BSF-R section of the Child Activity Booklet, the interviewer asked the parent respondent's opinion about the child's performance. The respondent's answers to these two questions have important implications for the child's BSF-R scores. If it was necessary to break off and complete the BSF-R in a second visit, these same questions were asked again to obtain the parent respondent's opinion about the child's performance during this second visit.

The first question asked the respondent whether the child's behavior during the BSF-R was typical, whether the child was as alert and active as usual, and whether the child was as happy or upset as usual. Based on the respondent's answer, the interviewer then rated the response on a 5-point scale ranging from "(1) Very atypical" (caregiver never sees this type of behavior) to "(5) Very typical" (caregiver always sees this type of behavior). If the child's behavior during the BSF-R was reported by the respondent as being in the atypical range (i.e., a rating of 1 or 2), then the child's BSF-R scores may underestimate the child's true level of functioning. On the other hand, if the child's performance was reported by the respondent as being in the typical range (i.e., a rating of 3, 4, or 5), then the child's testing behaviors were representative of the child's general level of functioning.

The second question asked whether the parent respondent thought the child had done as well as expected on the BSF-R or whether the child had done better or worse on similar types of activities. Again, the interviewer rated the respondent's answer on a 5-point scale ranging from "(1) Poor indicator of child's optimal performance" to "(5) Excellent, child never performs better." If the respondent reported that the child's performance was below his or her optimal level of functioning (i.e., a rating of 1 or 2), then the child's BSF-R scores may underestimate the child's current level of functioning. If, on the other hand, the respondent indicated that the child's performance was optimal or close to optimal (i.e., a rating of 3, 4, or 5), then the child's BSF-R scores are representative of the child's general level of functioning.

For more detailed information about the interviewer observation of child behavior items, please see the *User's Manual for the ECLS-B Longitudinal 9-Month–2-Year Data File and Electronic Codebook (NCES 2006-046)* (Nord et al. 2006).

Table 9-1. Interviewer-completed observations of child behavior in the 2-year BSF-R compared with those in the 9-month data collection: 2003–04

| 2-year<br>variable name | 2-year variable label | 9-month<br>variable name |
|---|---|---:|
| R2POSAFF | R2 CO035 CHILD DISPLAYS POSITIVE AFFECT | R1POSAFF |
| R2NEGAFF | R2 CO040 CHILD DISPLAYS NEGATIVE AFFECT | R1NEGAFF |
| R2ADAPT | R2 CO060 CHILD ADAPT CHANGE IN MATERIAL | R1ADAPT |
| R2INTRST | R2 CO065 CH SHOWS INTEREST IN MATERIAL | R1INTRST |
| R2ATNTSK | R2 CO080 CHILD PAYS ATTENTION TO TASKS | R1ATNTSK |
| R2PRSSTN | R2 CO085 CHILD PERSISTENT IN TASKS | † |
| R2FRFLNS | R2 CO095 CHILD DISPLAYS FEARFULNESS | † |
| R2FRSTRN | R2 CO100 CH DISPLAY FRUSTRATION IN TSKS | † |
| R2SOCIAL | R2 CO110 CH DISPLAYS SOCIAL ENGAGEMENT | R1SOCIAL |
| R2CHCOOP | R2 CO115 CHILD DISPLAYS COOPERATION | † |
| R2CNTLMV | R2 CO130 CH SHOWS CONTROL OF MOVEMENTS | R1CNTLMV |

†Not applicable. Item not administered during 9-month data collection.
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.

Table 9-2. Parent interview items on child behavior in the 2-year BSF-R compared with those in the 9-month data collection: 2003–04

| 2-year variable name | 2-year variable label (description) | 9-month variable name |
|---|---|---|
| C2BBHAV1 | C2 BSF CG RATE BEHAVIOR TYPICAL VISIT 1<br>(How typical was your child's behavior? Did {CHILD} play the way {he/she} usually does? Was {he/she} as happy or upset as usual? As alert and active as usual?) | C1BBHAV1 |
| C2BBHAV2 | C2 BSF CG RATE BEHAVIOR TYPICAL VISIT 2<br>(How typical was your child's behavior? Did {CHILD} play the way {he/she} usually does? Was {he/she} as happy or upset as usual? As alert and active as usual?) | C1BBHAV2 |
| C2PRFM1 | C2 BSF CG RATE PERFORMANCE VISIT 1<br>(Do you think {CHILD} did as well as {he/she} could? Have you seen {CHILD} do better or worse on the type of things we worked on?) | C1PRFM1 |
| C2PRFM2 | C2 BSF CG RATE PERFORMANCE VISIT 2<br>(Do you think {CHILD} did as well as {he/she} could? Have you seen {CHILD} do better or worse on the type of things we worked on?) | C1PRFM2 |

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.

### 9.1.1 Results From the 2-Year National Data Collection

To demonstrate the performance of these items in the ECLS-B 2-year national data collection, sample means and standard deviations were obtained, applying the longitudinal child weight W2C0 and omitting all cases with missing data. The descriptives for these items are presented in tables 9-3 and 9-4.

Table 9-3. Variable names, descriptions, average scores, and standard deviations for interviewer-completed observations about child behavior during the BSF-R in the 2-year data collection: 2003–04

| 2-year<br>variable name | Variable label | Number | Weighted<br>mean | Standard<br>deviation |
|---|---|---|---|---|
| R2POSAFF | R2 CO035 CHILD DISPLAYS POSITIVE AFFECT | 8,850 | 3.50 | 1.57 |
| R2NEGAFF | R2 CO040 CHILD DISPLAYS NEGATIVE AFFECT | 8,850 | 3.38 | 1.29 |
| R2ADAPT | R2 CO060 CHILD ADAPT CHANGE IN MATERIAL | 8,850 | 3.61 | 1.09 |
| R2INTRST | R2 CO065 CH SHOWS INTEREST IN MATERIAL | 8,850 | 3.50 | 0.94 |
| R2ATNTSK | R2 CO080 CHILD PAYS ATTENTION TO TASKS | 8,850 | 3.45 | 0.98 |
| R2PRSSTN | R2 CO085 CHILD PERSISTENT IN TASKS | 8,850 | 3.40 | 1.06 |
| R2FRFLNS | R2 CO095 CHILD DISPLAYS FEARFULNESS | 8,850 | 4.09 | 0.97 |
| R2FRSTRN | R2 CO100 CH DISPLAY FRUSTRATION IN TSKS | 8,850 | 3.74 | 0.99 |
| R2SOCIAL | R2 CO110 CH DISPLAYS SOCIAL ENGAGEMENT | 8,850 | 3.50 | 1.10 |
| R2CHCOOP | R2 CO115 CHILD DISPLAYS COOPERATION | 8,850 | 3.39 | 1.06 |
| R2CNTLMV | R2 CO130 CH SHOWS CONTROL OF MOVEMENTS | 8,850 | 4.30 | 0.72 |

NOTE: Sample sizes have been rounded to the nearest 50.
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.

Table 9-4. Variable names, descriptions, average scores, and standard deviations for parent interview items about child behavior during the BSF-R in the 2-year data collection: 2003–04

| 2-year<br>variable name | Variable label | Number | Weighted<br>mean | Standard<br>deviation |
|---|---|---|---|---|
| C2BBHAV1 | C2 BSF CG RATE BEHAVIOR TYPICAL VISIT 1 | 9,000 | 3.86 | 0.89 |
| C2BBHAV2 | C2 BSF CG RATE BEHAVIOR TYPICAL VISIT 2 | 100 | 3.39 | 1.21 |
| C2PRFM1 | C2 BSF CG RATE PERFORMANCE VISIT 1 | 9,000 | 3.44 | 1.01 |
| C2PRFM2 | C2 BSF CG RATE PERFORMANCE VISIT 2 | 100 | 2.87 | 1.26 |

NOTE: Sample sizes have been rounded to the nearest 50.
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.

When the BSID-II is administered in a clinical setting, the items on the BRS (in addition to other sources of information, such as a clinical interview with the parent) are used to confirm, call into question, or explain the child’s performance on the BSID-II. For example, a tired child or one who is ill may not perform up to expectations, and the child's scores on various BRS items may be used to explain his or her actual performance. Because only a small subset of BRS items was included in the ECLS-B, it was not possible to emulate their use in the full BRS. However, the child's behavior during the BSF-R should have some correlation with the BSF-R scores. Tables 9-5 and 9-6 present correlations between the BRS items used in the ECLS-B and children's scores on the BSF-R at 2 years. Please note that the BRS item ratings are unidirectional so that a higher score indicates either a greater frequency and intensity of positive behaviors (e.g., positive affect) or a lower frequency and intensity of negative behaviors (e.g., negative affect). Therefore, all the correlations in this table are positive. To obtain these correlations, the 2-year child weight W2C0 was applied and all cases with missing data were omitted.

Table 9-5. Correlations of major 2-year BSF-R mental and motor scores with the interviewer observation items, 2-year data collection: 2003–04

| | BSF-R score | | | |
|---|---|---|---|---|
| Interviewer observation item | X2MTLTSC (Mental *T* score) | X2MTLSCL (Mental scale score) | X2MTRTSC (Motor *T* score) | X2MTRSCL (Motor scale score) |
| R2 CO035 child displays positive affect | .37 | .38 | .29 | .32 |
| R2 CO040 child displays negative affect | .30 | .31 | .26 | .27 |
| R2 CO060 child adapts to change in materials | .40 | .41 | .25 | .27 |
| R2 CO065 child shows interest in materials | .52 | .53 | .34 | .37 |
| R2 CO080 child pays attention to tasks | .55 | .57 | .37 | .41 |
| R2 CO110 child displays social engagement | .30 | .30 | .27 | .28 |
| R2 CO085 child is persistent in tasks | .52 | .54 | .35 | .39 |
| R2 CO095 child displays fearfulness | .24 | .23 | .27 | .26 |
| R2 CO100 child displays frustration in tasks | .21 | .21 | .15 | .16 |
| R2 CO115 child displays cooperation | .51 | .52 | .36 | .39 |
| R2 CO130 child shows control of movements | .19 | .22 | .18 | .23 |

NOTE: The child weight W2C0 was applied. All correlations significant at $p < .0001$. $n = 8,550$ (rounded to the nearest 50).
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.

Table 9-6. Correlations of major 2-year BSF-R mental and motor scores with the two questions for parents, 2-year data collection: 2003–04

| | BSF-R score | | | |
|---|---|---|---|---|
| Parent interview item | X2MTLTSC<br>(Mental *T* score) | X2MTLSCL<br>(Mental scale score) | X2MTRTSC<br>(Motor *T* score) | X2MTRSCL<br>(Motor scale score) |
| C2 BSF caregiver rates behavior typical, visit 1 | .22 | .21 | .15 | .15 |
| C2 BSF caregiver rates performance, visit 1 | .31 | .32 | .25 | .27 |

NOTE: The child weight W2C0 was applied. All correlations significant at $p < .0001$. $n = 8{,}550$ (rounded to the nearest 50).
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.

### 9.1.2 Comparison of 2-Year Results With 9-Month Results

BRS items for the ECLS-B were selected to sample discrete behaviors that were developmentally appropriate for 2-year-olds and that were easily observed and rated by field interviewers. They were not selected with the intention of creating a subscale of BRS items. Therefore, in order to compare the results obtained at 2 years with those from 9 months, the correlations were obtained for those items included in both data collections. Table 9-7 presents these results.

### 9.1.3 Training on the Interviewer Observations of the Child’s Behavior

Training methods for the interviewer observation items at 2 years were the same as those used at 9 months, which had been successful. For the 2-year training, the videotapes were updated to include examples of toddler behaviors because the infant videotapes were no longer developmentally appropriate. The 11 observation items received direct attention at training with an item-by-item review of the Child Observations section of CAPI, accompanied by the videotape examples and a worksheet to monitor trainees’ understanding of the items. During a 2-hour session, trainees viewed videotapes of the target behaviors depicting both ends of the rating scale used to evaluate each item. For example, for the item “Child displays positive affect,” the trainees saw a video clip of a child broadly smiling and laughing and a video clip of a child who gave a fleeting and weak, but noticeable, smile. After review and discussion of each set of examples on the videotape, trainees viewed a “quiz videoclip” that they then

Table 9-7. Intercorrelations of 9-month and 2-year interviewer observations of child behavior items, 9-month and 2-year data collections: 2001–02 and 2003–04

| 2-year item | 9-month items | | | | | | |
|---|---|---|---|---|---|---|---|
| | R1POSAFF | R1NEGAFF | R1ADAPT | R1INTRST | R1ATNTSK | R1SOCIAL | R1CNTLMV |
| R2POSAFF | .18 | .05 | .07 | .11 | .12 | .15 | .11 |
| R2NEGAFF | .05 | .16 | .10 | .09 | .10 | .03 | .08 |
| R2ADAPT | .06 | .08 | .10 | .08 | .12 | .05 | .10 |
| R2INTRST | .13 | .07 | .08 | .16 | .16 | .13 | .14 |
| R2ATNTSK | .12 | .08 | .09 | .14 | .17 | .12 | .15 |
| R2PRSSTN | .11 | .08 | .10 | .14 | .16 | .11 | .16 |
| R2FRFLNS | .10 | .06 | .06 | .10 | .10 | .11 | .08 |
| R2FRSTRN | .08 | .07 | .09 | .11 | .12 | .07 | .11 |
| R2SOCIAL | .12 | .06 | .06 | .13 | .13 | .18 | .09 |
| R2CHCOOP | .10 | .08 | .09 | .11 | .15 | .10 | .13 |
| R2CNTLMV | .10 | .08 | .06 | .12 | .13 | .07 | .21 |

NOTE: The child weight W2C0 was applied to produce these statistics. All correlations significant at $p < .0001$. $n$= 8,750 (rounded to the nearest 50).

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) 9-month and 2-year data collections, 2001–02 and 2003–04.

rated on a worksheet collected at the end of the session. The training staff then reviewed these worksheets to identify any trainees having problems recognizing the behaviors. As with the physical measurements, the purpose here was to find interviewers who were having problems, rather than just to test the interviewers on their observations. Any interviewers having problems identifying the target behaviors were required to attend a help lab or otherwise receive further instruction from their trainers, depending on the extent of the problem. By the end of training, all interviewers had successfully completed this exercise.

## 9.2 Interviewer Observations of the Child’s Home Environment: HOME Items

The second instrument in the Child Observations consisted of a two-part set of items derived from the Short Form of the Home Observation for Measurement of the Environment (HOME) (Caldwell and Bradley 1979, 2001) and from the National Household Education Survey (NHES), also sponsored by the National Center for Education Statistics (NCES), to assess the quality of children’s environments. The

NHES is a large-scale household-based survey that obtains information about the educational activities of the U.S. population.

The HOME Short Form consists of 21 items, which would be too lengthy for the ECLS-B and, therefore, a subset of 12 items was selected and included in the 9-month and 2-year data collections. The full HOME is often used in academic and large-scale surveys to measure key aspects of children's environments, including the quality of parental interaction, the literacy environment, and the home environment. Several home environment items from NHES are similar to items from the HOME Short Form, with only some changes in wording and response categories. For further information about how the subset of 12 items was selected, please refer to the *ECLS-B Methodology Report for the Nine-Month Data Collection (2001–02), Volume 1: Psychometric Characteristics* (NCES 2005–100) (Andreassen and Fletcher 2005).

Eight of these 12 items obtained information about characteristics of the child's environment and were completed in the Child Observations section of CAPI by the interviewer on the basis of observations made during the home visit. The interviewer recorded whether or not specific environmental characteristics were observed during the home visit such as whether the parent spoke to the child (both spontaneously and in response to the child), whether the parent hugged (or kissed) the child or used physical means of managing the child's behavior (such as hitting or physical restraint), provided toys to the child, kept the child in view at all times and provided a safe environment. Response options included 1 = yes, 2 = no, and 3 = no opportunity to observe. These items are listed in table 9-8, along with their means and standard deviations.

The remaining four items were asked of the parent respondent as part of the parent interview and obtained information about the frequency with which the parent and child engaged in such activities as singing songs, telling stories, reading books and going on errands. Response options included: 1 = none; 2 = one or two (times a week); 3 = three to six (times a week); and 4 = everyday. These questions, similar to items in the HOME, were adopted from NHES with the intention of providing comparability between the ECLS-B and NHES. Although they are similar in content to items in the HOME, their wording and response categories are more consistent with NHES. These items can be found in table 9-9.

Table 9-8. Variable names, variable labels, and summary statistics for interviewer observations of home environment, 2-year data collection: 2003–04

| Child observation item | Variable label | Weighted percent yes |
|---|---|---|
| R2RSPKCH[1] | R2 CO165 RESP SPOKE SPONTANEOUS 2CHILD | 91.08 |
| R2IORSVB[1] | R2 CO170 RESP RESPONDED VERBALLY TO CHILD | 87.32 |
| R2IOCRSS[1] | R2 CO175 RESP CARESS/KISS/HUG CHILD | 88.18 |
| R2IORSHT | R2 CO180 RESP SLAPPED/SPANKED CHILD | 3.19 |
| R2IOINTF | R2 CO185 RESP INTERFERED WITH CH ACTION | 19.65 |
| R2IOPTYS[1] | R2 CO190 RESP PROVIDED TOYS TO CHILD | 74.68 |
| R2IOINVW[1] | R2 CO195 RESP KEPT CHILD IN VIEW | 89.57 |
| R2IOENVS[1] | R2 CO200 PLAY ENVIRONMENT WAS SAFE | 92.46 |

[1] These items were reverse coded.
NOTE: All data weighted using the parent weight, W2R0. *n* = 8,300 (rounded to the nearest 50).
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.

Table 9-9. Variable names, variable labels, and summary statistics for home environment parent interview items, 2-year data collection: 2003–04

| Parent interview item | Variable label | Weighted mean | Standard deviation |
|---|---|---|---|
| P2READBO | P2 HE075A HOW OFTEN YOU READ TO CHILD | 3.17 | 0.89 |
| P2TELLST | P2 HE075B HOW OFTN YOU TELL CH STORIES | 2.70 | 1.02 |
| P2SINGSO | P2 HE075C HOW OFTEN YOU ALL SING SONGS | 3.55 | 0.74 |
| P2ERRAND | P2 HE075D HOW OFTN TAKE CHILD ON ERRAND | 3.50 | 0.73 |

NOTE: All data weighted using the parent weight, W2R0. *n* = 8,300 (rounded to the nearest 50).
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.

After the interviewer completed the observations of the child's behavior during the BSF-R, the interviewer observational items from the HOME were then completed on the laptop computer. The four questions from NHES were completed during the parent interview, in the Home Environment section.

For further information about the two parts of this subset of items, please see the *User's Manual for the ECLS-B Longitudinal 9-Month–2-Year Data File and Electronic Codebook* (NCES 2006–046) (Nord et al. 2006).

### 9.2.1 Comparison of 2-Year Results With 9-Month Results

To compare the 2-year home items with their performance at 9 months, the scalability of the 2-year home environment items was first examined by obtaining Cronbach's alpha, which had also been obtained at 9 months. Cronbach's alpha, a measure of internal consistency, is .46 for this set of home environment items at 2 years, which is low and calls into question the advisability of scaling the items. This result is similar to the alpha value obtained at the 9-month data collection, which was .50. The alphas at both data collections suggest that the items do not have a conceptual coherence and are not scalable. Other alternatives should be considered.

One possibility is to investigate the factor structure. As at 9 months, a principal components factor analysis with varimax rotation was conducted on the 2-year home environment items and found four factors with Eigenvalues greater than 1.0. The results were similar to those obtained at 9 months, with some differences. At 2 years, the factor explaining the most variance can be characterized as the child's language environment and cognitive stimulation. It includes the following: R2RSPKCH (interviewer observed mother speak to child), which had a factor load of 0.76; R2IORSVB (parent responded to child's verbalization), which had a factor load of 0.77; R2IOCRSS (parent caressed child), which had a factor load of 0.63; and R2TOPTYS (parent provided toys to child), which had a factor load of 0.52. This had been the second factor at 9 months. The second factor at 2 years can be characterized as the parent's literacy-oriented activities with the child. It includes the following: P2READBO (frequency parent reads books to child), which had a factor load of 0.74; P2TELLST (frequency parent tells stories to child), which had a factor load of 0.75; and P2SINSO (frequency parent sings songs), which had a factor load of 0.70. Similar to results at 9 months, the item P2ERRAND (frequency parent takes child on errands) had only a small loading on this factor at 0.28. This factor had been the first factor at 9 months, so factors one and two switched places across the two data collections, although they retain the same items. The third factor at 2 years consisted of R2IORSHT (interviewer observed parent hit child) and R2IOINTF (interviewer observed parent physically interfere with child's actions), which had factor loadings of 0.75 and 0.77, respectively. Both of these items are indications of physical methods for managing the child's behavior. The fourth factor at 2 years consisted of the items R2IOENVS (child's play environment was safe) with a factor loading of 0.64, and R2IOINVW (parent kept child in view), which had a factor loading of 0.60. P2ERRAND (parent takes child on errands) had only a modest negative loading on this factor at -0.45. These last two factors had also switched places from 9 months, when physical management of the child's behavior was the fourth factor. The items within each factor, however, are the same. These results from 9 months to 2 years suggest that as children develop, parents adjust their parenting skills to meet the needs of the child, for example, as children become more verbal parents also increase the frequency of their verbal responding.

### 9.2.2 Training for HOME Observation and Parent Interview Items

The subset of HOME observation items have been widely used in several large-scale, nationally representative surveys, such as the National Longitudinal Study of Youth (NLSY '79) and the National Institute of Child Health and Human Development (NICHD) Study of Early Child Care. The same items were also used in the 9-month data collection of the ECLS-B and yielded satisfactory data. Therefore, it was known that the training procedures used for the 9-month data collection would also be successful for the 2-year data collection. The classroom training on these observation items included direct review of each item as presented on the laptop screen in the Child Observations section of CAPI, although specific examples of the target behaviors were updated to be developmentally appropriate for 2-year-olds. The training of field interviewers on the HOME observation items was described previously in the *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) Methodology Report for the Nine-Month Data Collection (2001–02), Volume 1: Psychometric Characteristics* (NCES 2005–100) (Andreassen and Fletcher 2005).

# 10. INDIRECT ASSESSMENTS OF THE CHILD IN THE PARENT INTERVIEW

The Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) 2-year parent interview also included three sets of questions that obtained indirect assessments of the child's behavior from the parent respondent: a set of questions about the age at which the child reached certain developmental milestones, a set of questions about the child's self-regulation ability and sensory integration, and a set of questions about the child's vocabulary and language ability. The following sections discuss the rationale for these sets of questions, summarize the items, and present descriptive statistics for key demographic groups.

## 10.1 Developmental Milestones

The first indirect assessment was a set of six questions that asked the parent respondent about the age at which the child first reached developmental milestones common to the 2-year age range. A developmental milestone is a set of functional skills, or age-specific abilities, that most children can do by a certain age, for example, taking first steps on own, saying first word. Although each developmental milestone has an age level at which it is typically reached, the actual age at which a normally developing child reaches that milestone can vary quite a bit. Developmental milestones serve as reference points or benchmarks that parents, health care professionals, psychologists, and teachers can use to help check how a child is developing.

It is commonly presumed that early attainment of milestones is associated with positive outcomes and that later achievement of developmental milestones is associated with poorer developmental status and child outcomes in later years. However, until the ECLS-B, no national norms had been available to support the association of early milestone achievement with subsequent positive outcomes. Nor is there any accessible empirical evidence to suggest that early or timely achievement of developmental milestones has any bearing on future developmental status, although ample evidence supports the association between late achievement of milestones (e.g., as is common in children with Down syndrome) and poorer developmental status in later years.

To identify key developmental milestones for the 23- to 25-month age range, the Minnesota Child Development Inventory (MN-CDI) (Ireton 1997) was reviewed as a source for items, as it had been for the 9-month data collection. For further information about the five developmental milestone items in the 9-month data collection, please refer to the *ECLS-B Methodology Report for the Nine-Month Data Collection (2001–02), Volume 1: Psychometric Characteristics* (NCES 2005–100) (Andreassen and Fletcher 2005). For obvious reasons, the milestones selected for 9 months are not appropriate at 2 years. Therefore, the age ranges in the MN-CDI manual were reviewed and key milestones that were age appropriate, easily understood, and salient to parents were selected for inclusion in the ECLS-B at 2 years. It was important that the milestones selected be particularly salient for parents who would be formulating their answers retrospectively. In addition, it was important that the response options be straightforward and not lead to embarrassment if the child had not yet reached certain milestones.

The results of field testing and supplementary review of the MN-CDI for redesign of the parent interview for 2 years determined that the following items most successfully obtained information about the age at which the child first performed the milestone:

- P2WLKSTR: How old was child in months when he/she started walking up stairs alone?
- P2FRSTWD: How old was child in months when he/she started saying words?
- P2TRNPGS: How old was child in months when he/she started turning the pages of a picture book, one at a time?
- P2DRKNB: How old was child in months when he/she started opening a door by turning the knob and pulling?
- P2PLYOH: How old was child in months when he/she started playing with other children, doing things with them (e.g., cars, dolls, building)?
- P2PLYOB: How old was child in months when he/she starting using an object as if it were something else (e.g., using a block for a phone, using a cardboard box for a car or a doll bed, using a napkin for a doll blanket)?

Table 10-1 presents the percentages of sample children who had reached each milestone at the time of the child assessment, as well as the percentages who had not yet reached each milestone. To obtain these weighted percentages, all cases with missing data were omitted. This information is presented in this table to highlight that not all children in the ECLS-B had reached all the milestones by the time of the home visit. Information about the ages at which children passed each milestone is presented in table 10-2. No comparable national norms are available for these milestones.

Table 10-1. Weighted percentages of children who have reached and have not yet reached developmental milestones by the time of assessment in the 2-year data collection, with item names and descriptions, 2-year data collection: 2003–04

| | | Reached milestone | | Not yet reached milestone | |
|---|---|---|---|---|---|
| Developmental milestone | Description of milestone | Number | Weighted percent | Number | Weighted percent |
| P2WLKSTR | First started walking upstairs alone | 9,000 | 93.89 | 800 | 6.11 |
| P2FRSTWD | Started saying first words | 9,650 | 99.46 | 100 | 0.54 |
| P2TRNPGS | Started turning pages of book one at a time | 9,250 | 96.02 | 450 | 3.98 |
| P2DRKNB | Started opening door by turning knob | 7,550 | 80.22 | 2,250 | 19.78 |
| P2PLYOH | Started playing with other children | 9,400 | 96.83 | 400 | 3.17 |
| P2PLYOB | Started using as object as if it were something else | 9,100 | 94.31 | 650 | 5.69 |

NOTE: The parent weight, W2R0, was used to obtain these statistics. Cell counts are unweighted to show the distribution in the ECLS-B 2-year data collection. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.
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.

Table 10-2 presents the average ages (weighted means and standard deviations) at which children reached each developmental milestone by the time of the home visit for the total sample and by demographic grouping variables. To obtain these statistics, all cases with missing data were omitted. The weighted means suggest that, for example, girls tend to start walking up stairs at slightly younger ages than boys, at 15.32 months of age for girls and 15.39 months of age for boys. Black children start to play with other children at an average of 12.02 months of age, whereas White children start to play with other children at an average of 14.20 months of age.

Table 10-2. Average age weighted means and standard deviations for developmental milestones, by key demographic characteristics, 2-year data collection: 2003–04

| Demographic | Developmental milestones | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Walked up stairs alone (P2WLKSTR) | | | Started saying first words (P2FRSTWD) | | | Turned pages of book (P2TRNPGS) | | | Opened door by turning knob (P2DRKNB) | | | Played with other children (P2PLYOH) | | | Played with object (P2PLYOB) | | |
| characteristics | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD |
| Total score | 9,000 | 15.36 | 3.75 | 9,650 | 11.11 | 3.86 | 9,250 | 13.72 | 4.12 | 7,550 | 18.44 | 3.94 | 9,400 | 13.55 | 4.53 | 9,100 | 15.30 | 4.15 |
| | | | | | | | | | | | | | | | | | | |
| Mother's race/ethnicity[1] | | | | | | | | | | | | | | | | | | |
| White | 4,100 | 15.95 | 3.62 | 4,500 | 11.09 | 3.81 | 4,350 | 14.01 | 4.10 | 3,400 | 19.07 | 3.61 | 4,350 | 14.14 | 4.55 | 4,250 | 15.92 | 3.97 |
| Black | 1,400 | 14.29 | 3.84 | 1,550 | 10.82 | 3.77 | 1,450 | 13.30 | 4.02 | 1,200 | 17.35 | 4.33 | 1,500 | 11.97 | 4.20 | 1,450 | 14.09 | 4.35 |
| Hispanic, race specified | 1,550 | 14.57 | 3.72 | 1,650 | 11.26 | 4.00 | 1,600 | 13.28 | 4.16 | 1,300 | 17.64 | 4.10 | 1,600 | 13.07 | 4.36 | 1,550 | 14.56 | 4.21 |
| Hispanic, no race specified | 50 | 13.77 | 2.88 | 50 | 12.30 | 4.95 | 50 | 12.77 | 4.08 | 50 | 16.65 | 4.22 | 50 | 12.33 | 3.95 | 50 | 13.02 | 2.38 |
| Asian | 1,150 | 14.86 | 3.74 | 1,200 | 11.43 | 3.81 | 1,150 | 13.70 | 4.23 | 1,050 | 18.11 | 4.00 | 1,150 | 14.05 | 4.35 | 1,150 | 15.36 | 3.96 |
| Native Hawaiian/ Pacific Islander | 50 | 15.48 | 3.87 | 50 | 10.82 | 3.04 | 50 | 12.79 | 4.18 | 50 | 18.07 | 3.91 | 50 | 13.29 | 4.89 | 50 | 14.76 | 4.16 |
| American Indian/ Alaska Native | 350 | 15.65 | 3.52 | 350 | 11.12 | 3.89 | 350 | 13.77 | 3.97 | 250 | 18.93 | 3.80 | 350 | 13.37 | 4.67 | 350 | 15.26 | 4.33 |
| More than 1 race | 250 | 16.06 | 3.81 | 250 | 11.55 | 4.09 | 250 | 13.87 | 4.21 | 200 | 19.10 | 4.22 | 250 | 12.73 | 5.24 | 250 | 15.24 | 4.55 |
| | | | | | | | | | | | | | | | | | | |
| Poverty status | | | | | | | | | | | | | | | | | | |
| Below poverty threshold | 2,000 | 14.54 | 3.70 | 2,150 | 10.89 | 4.10 | 2,000 | 13.71 | 4.20 | 1,650 | 17.92 | 4.28 | 2,100 | 12.78 | 4.49 | 2,050 | 14.50 | 4.28 |
| At or above poverty threshold | 7,000 | 15.58 | 3.73 | 7,500 | 11.18 | 3.79 | 7,250 | 13.73 | 4.10 | 5,900 | 18.58 | 3.83 | 7,300 | 13.76 | 4.52 | 7,100 | 15.53 | 4.09 |

See notes at end of table.

Table 10-2. Average age weighted means and standard deviations for developmental milestones, by key demographic characteristics, 2-year data collection: 2003–04—Continued

| Demographic characteristics | Developmental milestones | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Walked up stairs alone (P2WLKSTR) | | | Started saying first words (P2FRSTWD) | | | Turned pages of book (P2TRNPGS) | | | Opened door by turning knob (P2DRKNB) | | | Played with other children (P2PLYOH) | | | Played with object (P2PLYOB) | | |
| | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD |
| Child's race/ethnicity[1] | | | | | | | | | | | | | | | | | | |
| White | 3,750 | 15.96 | 3.61 | 4,100 | 11.10 | 3.83 | 3,950 | 13.99 | 4.11 | 3,050 | 19.08 | 3.62 | 4,000 | 14.21 | 4.59 | 3,850 | 15.95 | 3.96 |
| Black | 1,400 | 14.40 | 3.89 | 1,500 | 10.82 | 3.78 | 1,400 | 13.32 | 4.01 | 1,200 | 17.40 | 4.33 | 1,500 | 12.02 | 4.26 | 1,400 | 14.11 | 4.35 |
| Hispanic, race specified | 1,300 | 14.72 | 3.75 | 1,350 | 11.30 | 3.98 | 1,300 | 13.34 | 4.18 | 1,100 | 17.70 | 4.14 | 1,300 | 13.11 | 4.17 | 1,300 | 14.68 | 4.17 |
| Hispanic, no race specified | 550 | 14.57 | 3.62 | 600 | 11.13 | 3.93 | 600 | 13.36 | 4.09 | 450 | 17.84 | 4.00 | 600 | 12.95 | 4.53 | 550 | 14.42 | 4.13 |
| Asian | 950 | 14.73 | 3.71 | 1,000 | 11.21 | 3.76 | 1,000 | 13.58 | 4.27 | 900 | 18.00 | 4.06 | 1,000 | 13.75 | 4.31 | 950 | 15.22 | 3.95 |
| Native Hawaiian/ Pacific Islander | 50 | 15.60 | 3.78 | 50 | 10.24 | 2.96 | 50 | 14.38 | 5.29 | 50 | 17.25 | 4.21 | 50 | 10.99 | 3.20 | 50 | 14.93 | 4.61 |
| American Indian/ Alaska Native | 250 | 15.85 | 3.71 | 250 | 11.38 | 4.26 | 250 | 14.14 | 4.17 | 200 | 18.69 | 3.92 | 250 | 13.82 | 4.74 | 250 | 15.18 | 4.18 |
| More than 1 race | 700 | 15.51 | 3.81 | 700 | 11.39 | 3.94 | 700 | 13.80 | 4.04 | 550 | 18.68 | 3.83 | 700 | 13.18 | 4.53 | 650 | 15.29 | 4.44 |
| Child's sex | | | | | | | | | | | | | | | | | | |
| Male | 4,550 | 15.39 | 3.74 | 4,900 | 11.48 | 4.00 | 4,700 | 13.95 | 4.17 | 3,950 | 18.36 | 3.95 | 4,800 | 13.62 | 4.52 | 4,650 | 15.48 | 4.20 |
| Female | 4,400 | 15.32 | 3.76 | 4,750 | 10.73 | 3.67 | 4,550 | 13.49 | 4.05 | 3,600 | 18.53 | 3.93 | 4,600 | 13.47 | 4.54 | 4,450 | 15.15 | 4.10 |
| Child's age at assessment | | | | | | | | | | | | | | | | | | |
| 21 months and under | # | 14.55 | 2.69 | # | 8.04 | 4.24 | # | 14.54 | 3.94 | # | 15.10 | 1.56 | # | 11.77 | 2.11 | # | 14.44 | 3.96 |
| 22–23 months | 850 | 15.04 | 3.43 | 950 | 11.12 | 3.88 | 900 | 13.57 | 4.10 | 700 | 17.93 | 3.76 | 900 | 13.24 | 4.27 | 900 | 15.06 | 3.88 |
| 24–25 months | 6,800 | 15.40 | 3.70 | 7,300 | 11.07 | 3.79 | 6,950 | 13.74 | 4.08 | 5,650 | 18.39 | 3.82 | 7,100 | 13.56 | 4.48 | 6,850 | 15.33 | 4.09 |
| 26–27 months | 1,000 | 15.39 | 4.11 | 1,050 | 11.34 | 4.17 | 1,050 | 13.63 | 4.20 | 900 | 18.93 | 4.44 | 1,050 | 13.76 | 4.90 | 1,050 | 15.38 | 4.48 |
| 28 months and over | 300 | 15.40 | 4.40 | 350 | 11.38 | 4.20 | 300 | 14.16 | 4.84 | 300 | 19.27 | 4.81 | 300 | 13.76 | 5.20 | 300 | 15.26 | 5.15 |

See notes at end of table.

Table 10-2. Average age weighted means and standard deviations for developmental milestones, by key demographic characteristics, 2-year data collection: 2003–04—Continued

| | Developmental milestones | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Demographic characteristics | Walked up stairs alone (P2WLKSTR) | | | Started saying first words (P2FRSTWD) | | | Turned pages of book (P2TRNPGS) | | | Opened door by turning knob (P2DRKNB) | | | Played with other children (P2PLYOH) | | | Played with object (P2PLYOB) | | |
| | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD |
| Child's birth weight | | | | | | | | | | | | | | | | | | |
| Normal | 6,750 | 15.32 | 3.74 | 7,100 | 11.06 | 3.83 | 6,850 | 13.70 | 4.11 | 5,850 | 18.41 | 3.94 | 6,950 | 13.55 | 4.53 | 6,750 | 15.30 | 4.14 |
| Moderately low | 1,400 | 15.56 | 3.74 | 1,500 | 11.45 | 4.04 | 1,400 | 13.87 | 4.15 | 1,100 | 18.76 | 3.94 | 1,450 | 13.34 | 4.53 | 1,400 | 15.17 | 4.26 |
| Very low | 800 | 17.25 | 3.90 | 1,000 | 13.09 | 4.27 | 950 | 14.86 | 4.26 | 600 | 19.19 | 3.94 | 950 | 14.55 | 4.79 | 900 | 16.34 | 4.22 |
| Mother's age (in years) | | | | | | | | | | | | | | | | | | |
| 19 and under | 300 | 14.41 | 3.56 | 350 | 10.25 | 3.84 | 300 | 13.56 | 4.34 | 250 | 18.30 | 4.25 | 300 | 12.83 | 4.29 | 300 | 14.61 | 4.38 |
| 20–29 | 4,050 | 15.20 | 3.76 | 4,300 | 10.78 | 3.79 | 4,100 | 13.60 | 4.02 | 3,350 | 18.40 | 4.00 | 4,200 | 13.18 | 4.38 | 4,050 | 14.99 | 4.14 |
| 30–39 | 3,900 | 15.57 | 3.75 | 4,250 | 11.45 | 3.89 | 4,100 | 13.84 | 4.19 | 3,350 | 18.51 | 3.88 | 4,100 | 14.03 | 4.69 | 4,000 | 15.70 | 4.13 |
| 40 and over | 650 | 15.63 | 3.73 | 750 | 11.77 | 3.90 | 700 | 13.90 | 4.25 | 600 | 18.37 | 3.68 | 700 | 13.54 | 4.42 | 650 | 15.37 | 4.07 |
| Mother's education | | | | | | | | | | | | | | | | | | |
| 8th grade or below | 400 | 14.14 | 3.74 | 450 | 11.30 | 4.24 | 450 | 13.40 | 4.50 | 350 | 17.41 | 4.28 | 450 | 13.15 | 4.27 | 400 | 14.29 | 4.37 |
| 9–12th grades | 1,800 | 14.58 | 3.77 | 1,950 | 10.62 | 3.91 | 1,850 | 13.30 | 4.08 | 1,500 | 17.84 | 4.27 | 1,900 | 12.68 | 4.33 | 1,850 | 14.59 | 4.27 |
| High school diploma | 1,900 | 14.98 | 3.75 | 2,050 | 11.01 | 3.80 | 1,950 | 13.70 | 4.06 | 1,550 | 18.36 | 4.15 | 2,050 | 12.92 | 4.37 | 1,950 | 14.84 | 4.18 |
| Vocational/technical | 150 | 15.80 | 3.11 | 200 | 11.39 | 3.76 | 200 | 14.48 | 4.20 | 150 | 19.27 | 3.49 | 200 | 13.84 | 4.27 | 200 | 15.52 | 3.70 |
| Some college | 2,150 | 15.70 | 3.67 | 2,350 | 11.14 | 3.91 | 2,250 | 13.93 | 4.02 | 1,850 | 18.68 | 3.66 | 2,300 | 13.49 | 4.53 | 2,200 | 15.44 | 4.01 |
| Bachelor's degree | 1,450 | 16.21 | 3.63 | 1,550 | 11.63 | 3.76 | 1,550 | 13.96 | 4.22 | 1,250 | 18.96 | 3.64 | 1,500 | 14.86 | 4.63 | 1,500 | 16.36 | 3.99 |
| Graduate school (no degree) | 150 | 16.24 | 3.56 | 150 | 10.85 | 3.43 | 150 | 14.02 | 3.98 | 150 | 18.62 | 3.45 | 150 | 15.16 | 4.81 | 150 | 17.05 | 3.71 |
| Master's degree | 650 | 16.30 | 3.62 | 650 | 11.46 | 3.65 | 650 | 13.57 | 4.10 | 550 | 19.25 | 3.34 | 650 | 15.15 | 4.60 | 650 | 16.38 | 3.82 |
| Doctoral/professional degree | 200 | 16.62 | 3.47 | 250 | 11.65 | 3.27 | 250 | 14.39 | 4.16 | 200 | 18.29 | 3.25 | 200 | 15.78 | 3.95 | 200 | 16.18 | 3.69 |

# Rounds to zero.

[1] Race categories exclude Hispanic origin unless specified.

NOTE: The parent weight, W2R0, was used to obtain these statistics. Cell counts are unweighted to show the distribution in the ECLS-B 2-year data collection. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

## 10.2 Self-Regulatory Skills (Infant/Toddler Symptom Checklist)

The second set of indirect assessment questions were obtained from the *Infant/Toddler Symptom Checklist* (ITSC) (DeGangi et al. 1995). This checklist is a screener that was designed to be completed by parents and obtains information about children's self-regulatory behaviors and sensory integration. Sensory experiences include touch, movement, body awareness (proprioception), sight, sound, smell, and taste. Sensory integration is the process of distinguishing among these sensory experiences and is usually automatic and effortless. Children with sensory integration disorder may have difficulty achieving this integration or may achieve it only through extensive effort, attention, and frustration. For children with sensorimotor integration problems, sensory information, such as touch, sound, and movement, is misinterpreted for various neurophysiological reasons, for example, clothing tags at the back of the neckline are experiences as quite aversive to children with some types of sensorimotor integration problems. This neurophysiological misinterpretation, in turn, can lead to behavioral problems, difficulties with motor planning, motor coordination, and many other issues, including sustained attention, executive processing and, more generally, learning (Ayres 1979; Fisher, Murray, and Bundy 1991).

During the 9-month data collection, there were seven ITSC items in the parent computer-assisted personal interview (CAPI) instrument that were appropriate for this age range. There were two criteria for item selection. First, items were selected for their ability to identify children with sensorimotor and self-regulatory difficulties that are associated with attention or behavior problems or both, in the preschool years and later. In addition, items were chosen on the basis of the salience of the behavior to the parent, who would be recalling the information retrospectively; the parent had to be able to recognize the behavior clearly in order to report it accurately. The same criteria were used to select items for the 2-year data collection.

The ITSC was designed for the 7- to 30-month age range, and there are five age-appropriate versions (e.g., 7–9 months, 10–12 months). There are two age-appropriate versions that are relevant to the 2-year data collection: the 19- to 24-month version and the 25- to 30-month version. In the 19- to 24-month version, the full ITSC includes 23 items. In the 25- to 30-month version, the full ITSC includes 18 items. There are only seven items common to both the 19- to 24-month version and the 25- to 30-month version. The full complement of unique items would total 34 items, which would be too lengthy for the purpose of the ECLS-B. Therefore, a subset of seven items was selected for the ECLS-B Parent CAPI Instrument on the basis of the items' ability to identify children with regulatory disorders. The items selected for the 2-year data collection cover the domains of self-regulation, irritability, sleep difficulty, distractibility, and attending. Parents were asked how often their children were like the descriptions in each item. They indicated whether the child is "never" like this (0), "used to be" like this but is no longer (1), is "sometimes" like this (2), or is like this "most times" (3). The ECLS-B rating for these items could, therefore, range from 0 to 3. The items chosen include the following:

- P2FUSSY: Child is frequently irritable or fussy.
- P2WHMPR: Child goes easily from a whimper to an intense cry.
- P2UNBWT: Child is unable to wait for food or toys without crying or whining/falling apart.
- P2DSTRCT: Child is easily distractible or has fleeting attention.
- P2HLPSLP: Child needs a lot of help to fall asleep (e.g., rocking, long walks, stroking hair, car rides, etc.).
- P2TUNOUT: Child tunes out from activity and is difficult to re-engage.
- P2SFTFOC: Child can't shift focus easily from one project or activity to another.

These items were selected because they were identified in the ITSC manual as among those that are most successful at differentiating children with regulatory disorders from children without such disorders at this age. In addition, they were selected because these behaviors are salient to parents and easily reportable. The ITSC manual (DeGangi et al. 1995) presents age-specific summary tables for each item's ability to differentiate children who have regulatory disorders from children who do not have regulatory disorders. The tables for the 19- to 24-month version and the 25- to 30-month version were consulted because they were age appropriate. Table 10-3 presents a summary of the ITSC items selected from the two age-specific versions that were most successful at differentiating children with regulatory disorders from children without regulatory disorders. The column titled *t* value refers to the value of *t* that was obtained on a *t* test. Items were chosen that had a significant difference on the *t* test and that would be likely to be observed and reported by parents.

The analyst may want to conduct a factor analysis to explore the possibility of combining items to represent a particular construct in the 2-year ITSC data. Table 10-4 presents the 2-year ITSC item frequency distributions for the sample as a whole.

As described in the manual accompanying the ITSC, it is used to screen children who may be at risk and, therefore, would benefit from an intervention program. The manual presents age-appropriate cut-off scores by which to determine whether a child is at risk. However, the ECLS-B only uses about half of the items in the full ITSC. The analyst may want to consider prorating the summed scores and determining a prorated cut-off score by which to determine risk. Refer to the ITSC manual (DeGangi et al. 1995) for further information as well as to the *User's Manual for the ECLS-B Longitudinal 9-Month–2-Year Data File and Electronic Codebook* (NCES 2006–046) (Nord et al. 2006) for further information about how the ITSC scores can be used.

Table 10-3. Differentiation of regulatory disordered children by items from the 19- to 24-month and 25- to 30-month versions of the Infant/Toddler Symptom Checklist in the ECLS-B: 2003–04

| Variable name | ITSC item description | Mean score for children without regulatory disorders | Mean score for children with regulatory disorders | *t* value[1] | df |
|---|---|---|---|---|---|
| P2FUSSY | Child is frequently irritable, fussy | .10 | .93 | -5.46 * | 43 |
| P2WHMPR | Child goes easily from a whimper to an intense cry | .20 | .85 | -4.65 * | 41 |
| P2UNBWT | Child is unable to wait for food or toys without falling apart | .13 | 1.15 | -4.75 * | 41 |
| P2DSTRCT | Child is easily distractible or has fleeting attention | .20 | .85 | -2.63 * | 43 |
| P2HLPSLP | Child needs help to fall asleep | .35 | .79 | -2.10 * | 43 |
| P2TUNOUT | Child tunes out from activity, is difficult to reengage | .03 | .29 | -2.14 * | 43 |
| P2SFTFOC | Child can't shift focus easily from one project or activity to another | .03 | .50 | -2.94 * | 43 |

* $p < .05$.

[1] The *t* test statistic presented here indicates whether the item significantly differentiated children who have a regulatory disorder from children who do not have a regulatory disorder.

NOTE: These results are published in the manual for the ITSC (DeGangi et al. 1995) and are not from ECLS-B data. Values range from 0 (Child is never like this) to 3 (Child is like this most times).

SOURCE: DeGangi, G. A., Poisson, S., Sickel, R. Z., and Weiner, A. S. (1995). *Infant/Toddler Symptom Checklist: A Screening Tool for Parents*. San Antonio, Texas: Therapy Skill Builders, a division of The Psychological Corporation.

Table 10-4. Self-regulatory item frequency distributions for the total sample, 2-year data collection: 2003–04

| Variable name | Response option | Number | Weighted percent |
|---|---|---:|---:|
| P2FUSSY | (0) Never | 2,250 | 23.83 |
| | (1) Used to be | 600 | 5.44 |
| | (2) Sometimes | 6,200 | 63.06 |
| | (3) Most times | 800 | 7.68 |
| P2WHMPR | (0) Never | 3,800 | 41.76 |
| | (1) Used to be | 750 | 7.06 |
| | (2) Sometimes | 4,200 | 40.54 |
| | (3) Most times | 1,100 | 10.63 |
| P2UNBWT | (0) Never | 2,550 | 27.46 |
| | (1) Used to be | 700 | 7.02 |
| | (2) Sometimes | 4,850 | 49.60 |
| | (3) Most times | 1,700 | 15.92 |
| P2DSTRCT | (0) Never | 2,500 | 25.87 |
| | (1) Used to be | 550 | 5.44 |
| | (2) Sometimes | 4,950 | 50.41 |
| | (3) Most times | 1,800 | 18.28 |
| P2HLPSLP | (0) Never | 5,650 | 58.11 |
| | (1) Used to be | 950 | 9.45 |
| | (2) Sometimes | 1,900 | 19.31 |
| | (3) Most times | 1,350 | 13.13 |
| P2TUNOUT | (0) Never | 4,750 | 50.12 |
| | (1) Used to be | 450 | 4.47 |
| | (2) Sometimes | 3,950 | 39.31 |
| | (3) Most times | 650 | 5.89 |
| P2SFTFOC | (0) Never | 4,900 | 52.96 |
| | (1) Used to be | 450 | 4.89 |
| | (2) Sometimes | 3,700 | 35.46 |
| | (3) Most times | 750 | 6.70 |

NOTE: The parent weight, W2R0, was used to obtain these statistics. Cell counts are unweighted to show the distribution in the ECLS-B 2-year data collection. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

To examine how the items evaluating children's self-regulatory skills performed during the 2-year national data collection, the weighted means and standard deviations of each item were obtained for the total sample and for the key demographic variables. To obtain these statistics, the parent weight, W2R0, was applied and all cases with missing data were omitted. These statistics are presented in table 10-5. The weighted means demonstrate that there is variability across the items as well as across different demographic variables. For example, children born with very low birth weight (less than or equal to 1,500 grams) are more easily distractible (P2DSTRCT) than children born at greater than 1,500 grams, with an average of 1.80 versus 1.60 for children born at normal birth weight. In addition, children living below poverty level tend have higher average scores on all but one (P2HLPSLP) of the self-regulation items than those at or above poverty threshold.

## 10.3 Toddler Vocabulary

The acquisition of language is such an important developmental milestone that it deserves a measurement tool of its own. In fact, the transition from preverbal to verbal communication is so important that it marks the boundary between (preverbal) infancy and (verbal) toddlerhood. Children learn language at different rates, some early and quickly, some later and more slowly, so that the range of words acquired by a certain age is broad, with estimates ranging from 10 words acquired by 13 to 15 months and 50 words acquired by 10 to 24 months (Nelson 1973; Fenson et al. 1994). As reported on the Child Language Data Exchange System website (http://childes.psy.cmu.edu/), from 24 to 36 months language acquisition is rapid and accelerates steeply, so that a plausible estimate would be an average of 10 new words a day during the preschool and early school years.

Children learn language at different rates. Diary studies (e.g., Nelson 1973) and studies of the language environment in the home (e.g., dinner conversation studies, Beals and Snow 1994) have shown that language acquisition can be influenced by social factors. Therefore, the 2-year data collection of the ECLS-B is well served by the inclusion of a measure of the child's language acquisition and word use. A measure of children's language acquisition enables analysts to examine the variables in children's environments that contribute to higher rates of word learning, which, in turn, is presumed to be related to children's subsequent adjustment to and achievement in the early school years.

Table 10-5. Weighted means and standard deviations for children's self-regulatory behaviors, by key demographic characteristics, 2-year data collection: 2003–04

| Demographic characteristics | Self-regulatory behaviors | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Frequency irritability (P2FUSSY) | | | Goes easily from whimper to intense cry (P2WHMPR) | | | Unable to wait for food or toys (P2UNBWT) | | | Easily distractible/ fleeting attention (P2DSTRCT) | | | Needs help to fall asleep (P2HLPSLP) | | | Tunes out from activity (P2TUNOUT) | | | Can't shift focus easily (2SFTFOC) | | |
| | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD |
| Total score | 9,850 | 1.55 | 0.94 | 9,850 | 1.20 | 1.10 | 9,850 | 1.54 | 1.06 | 9,850 | 1.61 | 1.06 | 9,850 | 0.87 | 1.13 | 9,800 | 1.01 | 1.06 | 9,800 | 0.96 | 1.07 |
| Mother's race/ethnicity[1] | | | | | | | | | | | | | | | | | | | | | |
| White | 4,600 | 1.54 | 0.90 | 4,600 | 1.11 | 1.09 | 4,600 | 1.52 | 1.03 | 4,600 | 1.56 | 1.05 | 4,600 | 0.92 | 1.14 | 4,600 | 0.92 | 1.04 | 4,600 | 0.82 | 1.03 |
| Black | 1,550 | 1.70 | 0.90 | 1,550 | 1.40 | 1.09 | 1,550 | 1.74 | 1.03 | 1,550 | 1.76 | 1.03 | 1,550 | 0.80 | 1.09 | 1,550 | 1.16 | 1.08 | 1,550 | 1.14 | 1.08 |
| Hispanic, race specified | 1,650 | 1.45 | 1.02 | 1,600 | 1.28 | 1.12 | 1,650 | 1.49 | 1.11 | 1,650 | 1.64 | 1.08 | 1,650 | 0.77 | 1.11 | 1,650 | 1.14 | 1.11 | 1,650 | 1.17 | 1.12 |
| Hispanic, no race specified | 50 | 1.46 | 1.12 | 50 | 0.98 | 1.10 | 50 | 1.16 | 1.23 | 50 | 1.54 | 1.02 | 50 | 0.80 | 1.13 | 50 | 1.00 | 1.13 | 50 | 1.26 | 1.21 |
| Asian | 1,200 | 1.47 | 0.95 | 1,200 | 1.32 | 1.04 | 1,200 | 1.51 | 1.05 | 1,200 | 1.53 | 1.06 | 1,200 | 1.10 | 1.23 | 1,200 | 1.12 | 1.04 | 1,200 | 1.15 | 1.08 |
| Native Hawaiian/ Pacific Islander | 50 | 1.54 | 0.89 | 50 | 1.73 | 0.88 | 50 | 1.19 | 1.04 | 50 | 1.83 | 0.81 | 50 | 0.80 | 1.00 | 50 | 1.23 | 0.99 | 50 | 1.40 | 1.01 |
| American Indian/ Alaska Native | 350 | 1.76 | 0.82 | 350 | 1.34 | 1.11 | 350 | 1.64 | 1.05 | 350 | 1.65 | 1.08 | 350 | 0.90 | 1.08 | 350 | 0.89 | 1.06 | 350 | 1.00 | 1.05 |
| More than 1 race | 250 | 1.57 | 0.95 | 250 | 1.32 | 1.05 | 250 | 1.50 | 1.12 | 250 | 1.62 | 1.04 | 250 | 0.89 | 1.10 | 250 | 0.98 | 1.01 | 250 | 0.84 | 1.04 |
| Poverty status | | | | | | | | | | | | | | | | | | | | | |
| Below poverty threshold | 2,200 | 1.63 | 0.96 | 2,200 | 1.40 | 1.11 | 2,200 | 1.63 | 1.10 | 2,200 | 1.71 | 1.09 | 2,200 | 0.84 | 1.11 | 2,200 | 1.13 | 1.10 | 2,200 | 1.16 | 1.10 |
| At or above poverty threshold | 7,650 | 1.52 | 0.93 | 7,650 | 1.14 | 1.09 | 7,650 | 1.52 | 1.04 | 7,650 | 1.58 | 1.05 | 7,650 | 0.88 | 1.14 | 7,600 | 0.98 | 1.05 | 7,650 | 0.90 | 1.06 |

See notes at end of table.

Table 10-5. Weighted means and standard deviations for children's self-regulatory behaviors, by key demographic characteristics, 2-year data collection: 2003–04—Continued

| | Self-regulatory behaviors | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Demographic characteristics | Frequency irritability (P2FUSSY) | | | Goes easily from whimper to intense cry (P2WHMPR) | | | Unable to wait for food or toys (P2UNBWT) | | | Easily distractible/ fleeting attention (P2DSTRCT) | | | Needs help to fall asleep (P2HLPSLP) | | | Tunes out from activity (P2TUNOUT) | | | Can't shift focus easily (2SFTFOC) | | |
| | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD |
| Child's race/ethnicity[1] | | | | | | | | | | | | | | | | | | | | | |
| White | 4,200 | 1.54 | 0.90 | 4,200 | 1.10 | 1.09 | 4,200 | 1.52 | 1.03 | 4,200 | 1.56 | 1.05 | 4,200 | 0.91 | 1.14 | 4,200 | 0.91 | 1.03 | 4,200 | 0.81 | 1.03 |
| Black | 1,550 | 1.70 | 0.91 | 1,550 | 1.40 | 1.09 | 1,550 | 1.73 | 1.04 | 1,550 | 1.75 | 1.03 | 1,550 | 0.81 | 1.10 | 1,550 | 1.17 | 1.08 | 1,550 | 1.14 | 1.08 |
| Hispanic, race specified | 1,350 | 1.45 | 0.99 | 1,350 | 1.25 | 1.11 | 1,350 | 1.47 | 1.10 | 1,350 | 1.65 | 1.07 | 1,350 | 0.81 | 1.11 | 1,350 | 1.08 | 1.09 | 1,350 | 1.12 | 1.10 |
| Hispanic, no race specified | 600 | 1.53 | 1.04 | 600 | 1.28 | 1.13 | 600 | 1.48 | 1.13 | 600 | 1.65 | 1.09 | 600 | 0.79 | 1.14 | 600 | 1.22 | 1.12 | 600 | 1.21 | 1.14 |
| Asian | 1,050 | 1.45 | 0.95 | 1,050 | 1.37 | 1.03 | 1,050 | 1.49 | 1.07 | 1,050 | 1.51 | 1.07 | 1,050 | 1.08 | 1.23 | 1,050 | 1.14 | 1.04 | 1,050 | 1.19 | 1.08 |
| Native Hawaiian/ Pacific Islander | 50 | 1.67 | 0.88 | 50 | 1.36 | 1.11 | 50 | 1.26 | 1.19 | 50 | 1.65 | 0.96 | 50 | 1.02 | 1.23 | 50 | 1.67 | 0.98 | 50 | 1.22 | 1.04 |
| American Indian/ Alaska Native | 250 | 1.73 | 0.86 | 250 | 1.51 | 1.04 | 250 | 1.57 | 1.12 | 250 | 1.73 | 1.10 | 250 | 0.84 | 1.07 | 250 | 1.07 | 1.12 | 250 | 0.98 | 1.04 |
| More than 1 race | 750 | 1.56 | 0.94 | 750 | 1.25 | 1.07 | 750 | 1.59 | 1.06 | 750 | 1.59 | 1.06 | 750 | 0.90 | 1.16 | 750 | 0.98 | 1.04 | 750 | 0.92 | 1.04 |
| Child's sex | | | | | | | | | | | | | | | | | | | | | |
| Male | 5,050 | 1.56 | 0.93 | 5,050 | 1.19 | 1.10 | 5,050 | 1.59 | 1.05 | 5,000 | 1.68 | 1.05 | 5,000 | 0.87 | 1.13 | 5,000 | 1.06 | 1.07 | 5,000 | 0.99 | 1.08 |
| Female | 4,800 | 1.53 | 0.94 | 4,800 | 1.21 | 1.10 | 4,800 | 1.49 | 1.06 | 4,800 | 1.54 | 1.06 | 4,800 | 0.88 | 1.13 | 4,800 | 0.96 | 1.05 | 4,800 | 0.92 | 1.06 |
| Child's age at assessment | | | | | | | | | | | | | | | | | | | | | |
| 21 months and under | # | 2.04 | 0.62 | # | 1.54 | 0.97 | # | 2.31 | 0.63 | # | 2.10 | 0.30 | # | 1.73 | 1.18 | # | 1.47 | 0.93 | # | 1.78 | 1.11 |
| 22–23 months | 950 | 1.49 | 0.96 | 950 | 1.21 | 1.10 | 950 | 1.61 | 1.06 | 950 | 1.62 | 1.06 | 950 | 0.83 | 1.12 | 950 | 1.09 | 1.07 | 950 | 1.03 | 1.08 |
| 24–25 months | 7,400 | 1.57 | 0.92 | 7,400 | 1.20 | 1.09 | 7,400 | 1.54 | 1.05 | 7,400 | 1.63 | 1.05 | 7,400 | 0.89 | 1.14 | 7,400 | 1.00 | 1.06 | 7,400 | 0.95 | 1.07 |
| 26–27 months | 1,100 | 1.47 | 0.98 | 1,100 | 1.22 | 1.13 | 1,100 | 1.50 | 1.10 | 1,100 | 1.48 | 1.10 | 1,100 | 0.82 | 1.09 | 1,100 | 1.00 | 1.06 | 1,100 | 0.94 | 1.05 |
| 28 months and over | 350 | 1.45 | 0.99 | 350 | 1.15 | 1.08 | 350 | 1.40 | 1.04 | 350 | 1.48 | 1.06 | 350 | 0.84 | 1.12 | 350 | 0.98 | 1.06 | 350 | 0.90 | 1.04 |

See notes at end of table.

Table 10-5. Weighted means and standard deviations for children's self-regulatory behaviors, by key demographic characteristics, 2-year data collection: 2003–04—Continued

| | Self-regulatory behaviors | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Demographic characteristics | Frequency irritability (P2FUSSY) | | | Goes easily from whimper to intense cry (P2WHMPR) | | | Unable to wait for food or toys (P2UNBWT) | | | Easily distractible/ fleeting attention (P2DSTRCT) | | | Needs help to fall asleep (P2HLPSLP) | | | Tunes out from activity (P2TUNOUT) | | | Can't shift focus easily (2SFTFOC) | | |
| | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD |
| Child's birth weight | | | | | | | | | | | | | | | | | | | | | |
| Normal | 7,200 | 1.54 | 0.94 | 7,200 | 1.19 | 1.10 | 7,200 | 1.53 | 1.05 | 7,200 | 1.60 | 1.06 | 7,200 | 0.87 | 1.13 | 7,200 | 1.00 | 1.06 | 7,200 | 0.95 | 1.07 |
| Moderately low | 1,500 | 1.62 | 0.92 | 1,500 | 1.40 | 1.10 | 1,500 | 1.63 | 1.08 | 1,500 | 1.69 | 1.05 | 1,500 | 0.89 | 1.14 | 1,500 | 1.15 | 1.09 | 1,500 | 1.10 | 1.10 |
| Very low | 1,050 | 1.58 | 0.97 | 1,050 | 1.30 | 1.12 | 1,050 | 1.67 | 1.07 | 1,050 | 1.80 | 1.07 | 1,050 | 0.89 | 1.14 | 1,050 | 1.18 | 1.13 | 1,050 | 1.20 | 1.14 |
| Mother's age (in years) | | | | | | | | | | | | | | | | | | | | | |
| 19 and under | 350 | 1.62 | 0.92 | 350 | 1.41 | 1.09 | 350 | 1.66 | 1.05 | 350 | 1.96 | 1.01 | 350 | 0.95 | 1.17 | 350 | 1.15 | 1.10 | 350 | 1.02 | 1.13 |
| 20–29 | 4,400 | 1.62 | 0.93 | 4,400 | 1.26 | 1.10 | 4,400 | 1.62 | 1.05 | 4,400 | 1.70 | 1.05 | 4,400 | 0.84 | 1.12 | 4,400 | 1.09 | 1.08 | 4,400 | 1.02 | 1.08 |
| 30–39 | 4,300 | 1.47 | 0.94 | 4,300 | 1.13 | 1.09 | 4,300 | 1.47 | 1.06 | 4,300 | 1.51 | 1.06 | 4,300 | 0.87 | 1.13 | 4,300 | 0.92 | 1.04 | 4,300 | 0.89 | 1.06 |
| 40 and over | 750 | 1.51 | 0.93 | 750 | 1.12 | 1.09 | 750 | 1.39 | 1.04 | 750 | 1.48 | 1.05 | 750 | 1.09 | 1.22 | 750 | 0.99 | 1.06 | 750 | 0.96 | 1.08 |
| Mother's education | | | | | | | | | | | | | | | | | | | | | |
| 8th grade or below | 450 | 1.40 | 1.09 | 450 | 1.43 | 1.12 | 450 | 1.49 | 1.15 | 450 | 1.73 | 1.08 | 450 | 0.73 | 1.11 | 450 | 1.25 | 1.13 | 450 | 1.19 | 1.14 |
| 9–12th grades | 2,000 | 1.68 | 0.94 | 2,000 | 1.40 | 1.12 | 2,000 | 1.65 | 1.10 | 2,000 | 1.72 | 1.09 | 2,000 | 0.84 | 1.13 | 2,000 | 1.15 | 1.10 | 2,000 | 1.15 | 1.12 |
| High school diploma | 2,100 | 1.61 | 0.91 | 2,100 | 1.25 | 1.09 | 2,100 | 1.60 | 1.04 | 2,100 | 1.67 | 1.04 | 2,100 | 0.82 | 1.11 | 2,100 | 1.08 | 1.07 | 2,100 | 1.04 | 1.07 |
| Voc./technical | 200 | 1.57 | 0.87 | 200 | 1.04 | 1.07 | 200 | 1.43 | 1.07 | 200 | 1.69 | 0.93 | 200 | 0.93 | 1.19 | 200 | 1.11 | 1.07 | 200 | 0.81 | 1.03 |
| Some college | 2,350 | 1.47 | 0.94 | 2,350 | 1.11 | 1.09 | 2,350 | 1.51 | 1.05 | 2,350 | 1.57 | 1.06 | 2,350 | 0.91 | 1.15 | 2,350 | 0.94 | 1.04 | 2,350 | 0.86 | 1.03 |
| Bachelor's degree | 1,600 | 1.49 | 0.91 | 1,600 | 1.09 | 1.06 | 1,600 | 1.51 | 0.99 | 1,600 | 1.55 | 1.02 | 1,600 | 0.93 | 1.14 | 1,600 | 0.91 | 1.03 | 1,600 | 0.83 | 1.03 |
| Graduate school (no degree) | 150 | 1.57 | 0.85 | 150 | 1.13 | 1.04 | 150 | 1.43 | 0.98 | 150 | 1.51 | 1.07 | 150 | 0.93 | 1.10 | 150 | 0.88 | 1.01 | 150 | 0.84 | 1.00 |
| Master's degree | 700 | 1.41 | 0.90 | 700 | 0.86 | 1.03 | 700 | 1.33 | 1.00 | 700 | 1.30 | 1.04 | 700 | 1.03 | 1.13 | 700 | 0.73 | 0.95 | 700 | 0.69 | 0.99 |
| Doctoral/prof. degree | 250 | 1.41 | 0.90 | 250 | 1.10 | 1.00 | 250 | 1.34 | 0.99 | 250 | 1.34 | 1.06 | 250 | 0.83 | 1.10 | 250 | 0.68 | 0.98 | 250 | 0.55 | 0.94 |

# Rounds to zero.

[1] Race categories exclude Hispanic origin unless specified.

NOTE: The parent weight, W2R0, was used to obtain these statistics. Cell counts are unweighted to show the distribution in the ECLS-B 2-year data collection. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

There are a few items in the 2-year Bayley Short Form–Research Edition (BSF-R) that assess children's language use, and scores for these items are represented in children's BSF-R mental scale scores. To supplement information about children's language acquisition, a parent report of children's word use and grammatical constructions is included in the Parent CAPI Instrument. Ideally, the MacArthur Communicative Development Inventory (M-CDI) (Fenson et al. 1994) would have been the best available parent report measure. However, the M-CDI checklist contains well over 400 words, which would be too burdensome and time consuming to incorporate into the ECLS-B home visit.

For this reason, one of the co-authors of the M-CDI, Dr. Philip Dale of the University of Missouri, was contacted and agreed to develop a list of 50 words typically known and said by children in the target age range, as well as a set of items that obtain information about children's syntax use. These 50 words were then incorporated into the Parent CAPI Instrument to be read to the parent by the interviewer. This eliminated any difficulties filling out a checklist by respondents for whom English is not the primary language. Dr. Dale also recommended the set of supplementary items about children's language use, such as the use of irregular plurals and the use of irregular past tense. Therefore, the parent was asked whether the child: (1) could say the target words; (2) could combine words to make phrases or sentences; (3) could add "s" to make nouns plural; (4) could add " 's" to denote ownership; (5) could add "ing" to talk about activities in the present tense; and, (5) could add "ed" to words to talk about the past. In addition, Dr. Dale also provided an equivalent Spanish checklist, which included Spanish words that were appropriate for this age range and that were of approximately the equivalent level of difficulty as the English checklist. For further information about the variable names for these word checklist items and the supplementary syntax items, please refer to the *User's Manual for the ECLS-B Longitudinal 9-Month–2-Year Data File and Electronic Codebook* (NCES 2006–046) (Nord et al. 2006). The frequency distribution and percentages of sample children who had these words and did not have these words in their vocabulary are summarized in table 10-6. Table 10-7 summarizes children's syntax items.

There are no norms based on a nationally representative sample against which to compare the ECLS-B sample children. The best available measure that obtains comparable information is the M-CDI (Fenson et al. 1994), which includes a parent report checklist of the words children can say and the words children can understand. Interested analysts may want to refer to the monograph that describes the development of the M-CDI (Fenson et al. 1994) to see how scores were obtained on this measure. The results from the ECLS-B cannot be compared directly with the norms obtained for the M-CDI because the ECLS-B included only 50 words, whereas the age-appropriate toddler version of the M-CDI includes approximately 400 words from 19 categories.

Table 10-6 presents the (weighted) frequency distribution of children who said each word and did not say each word, as reported by the parent respondent. These distributions show that there is variability in children's vocabularies, ranging from a high of 96.22 percent of children who say "no" (only 3.78 percent do not yet say "no") and a low of 19.00 percent who can say "beside." To obtain these statistics, all cases with missing data were omitted and the parent weight, W2R0, was used.

Table 10-7 presents the frequency distribution of parents' reports of children's use of language rules (syntax) for the total sample. To obtain these statistics, the parent weight, W2R0, was used and all cases with missing data were omitted. However, the cell counts are unweighted in order to demonstrate the distribution in the ECLS-B 2-year data collection.

To examine how the items evaluating the child's vocabulary performed during the 2-year national data collection, the total score for this set of items was obtained by simply summing the number of words the child was able to say. The weighted means and standard deviations for children's total vocabulary score were obtained for the total sample and for the key demographic variables. These results are presented in table 10-8. To obtain these statistics, the parent weight, W2R0, was applied and all cases with missing data were omitted.

Table 10-9 shows the weighted means for the items that assess children's use of language rules (syntax) for the total sample and by demographic variables. To obtain these statistics, the parent weight, W2R0, was used and all cases with missing data were omitted. The cell counts, however, are unweighted in order to demonstrate the distribution in the ECLS-B 2-year data collection. For example, child age at the time of the assessment is associated with use of language rules: children 28 months of age and older score higher on many of these variables than do children who were assessed at 24 to 25 months.

Table 10-6. Frequency distribution of children's vocabulary items in the Parent CAPI Instrument, 2-year data collection: 2003–04

| Does child say… | Variable name | Says word |  | Does not say word |  |
|---|---|---|---|---|---|
|  |  | Number | Weighted percent | Number | Weighted percent |
| MEOW | P2SYMEOW | 7,600 | 81.12 | 2,250 | 18.88 |
| SHOE | P2SYSHOE | 8,550 | 88.44 | 1,300 | 11.56 |
| MOMMY | P2SYMMY | 9,300 | 92.91 | 550 | 7.09 |
| FAST | P2SYFAST | 4,050 | 45.74 | 5,750 | 54.26 |
| UHOH | P2SYUHOH | 9,000 | 91.14 | 850 | 8.86 |
| CHIN | P2SYCHIN | 4,350 | 47.05 | 5,500 | 52.95 |
| BYE | P2SYBYE | 9,350 | 95.70 | 500 | 4.30 |
| HOT | P2SYHOT | 8,450 | 87.82 | 1,350 | 12.18 |
| BEAR | P2SYBEAR | 6,300 | 66.08 | 3,500 | 33.92 |
| HAND | P2SYHAND | 7,300 | 77.17 | 2,550 | 22.83 |
| NO | P2SYNO | 9,400 | 96.22 | 450 | 3.78 |
| TINY | P2SYTINY | 2,400 | 25.76 | 7,400 | 74.24 |
| CAT | P2SYCAT | 8,100 | 85.63 | 1,700 | 14.37 |
| BROOM | P2SYBRM | 4,450 | 48.02 | 5,350 | 51.98 |
| THANK YOU | P2SYTHNK | 8,400 | 87.70 | 1,450 | 12.30 |
| AFTER | P2SYAFTR | 2,000 | 24.13 | 7,800 | 75.87 |
| DUCK | P2SYDUCK | 6,850 | 71.67 | 3,000 | 28.33 |
| MOP | P2SYMOP | 4,100 | 45.19 | 5,700 | 54.81 |
| CHASE | P2SYCHS | 2,650 | 30.61 | 7,150 | 69.39 |
| TONIGHT | P2SYTNGT | 3,650 | 39.15 | 6,150 | 60.85 |
| AIRPLANE | P2SYARPL | 6,450 | 70.18 | 3,350 | 29.82 |
| TRASH | P2SYTRSH | 6,150 | 66.01 | 3,650 | 33.99 |
| FISH | P2SYFSH | 4,500 | 47.72 | 5,350 | 52.28 |
| THEM | P2SYTHEM | 2,500 | 28.73 | 7,300 | 71.27 |
| CAR | P2SYCAR | 8,550 | 88.25 | 1,300 | 11.75 |
| TOWEL | P2SYTWL | 4,950 | 52.88 | 4,900 | 47.12 |
| HUG | P2SYHUG | 7,050 | 73.17 | 2,800 | 26.83 |
| US | P2SYUS | 3,550 | 40.20 | 6,300 | 59.80 |
| BOOK | P2SYBOOK | 8,650 | 89.96 | 1,200 | 10.04 |
| BEDROOM | P2SYBDRM | 4,400 | 48.18 | 5,450 | 51.82 |
| LIKE | P2SYLK | 4,600 | 51.22 | 5,200 | 48.78 |
| BESIDE | P2SYBSD | 1,500 | 19.00 | 8,300 | 81.00 |
| APPLESAUCE | P2SYAPSC | 3,550 | 41.67 | 6,250 | 58.33 |
| OVEN | P2SYOVEN | 3,900 | 44.03 | 5,900 | 55.97 |
| RIP | P2SYRIP | 2,200 | 24.67 | 7,650 | 75.33 |

See notes at end of table.

Table 10-6. Frequency distribution of children's vocabulary items in the Parent CAPI Instrument, 2-year data collection: 2003–04—Continued

| Does child say… | Variable name | Says word | | Does not say word | |
|---|---|---|---|---|---|
| | | Number | Weighted percent | Number | Weighted percent |
| UNDER | P2SYUNDR | 3,800 | 43.51 | 6,000 | 56.49 |
| COKE | P2SYCOKE | 4,700 | 52.47 | 5,100 | 47.53 |
| FLAG | P2SYFLAG | 3,400 | 38.53 | 6,450 | 61.47 |
| TASTE | P2SYTST | 3,650 | 41.14 | 6,150 | 58.86 |
| MUCH | P2SYMUCH | 3,100 | 34.20 | 6,750 | 65.80 |
| JUICE | P2SYJUCE | 8,400 | 85.28 | 1,450 | 14.72 |
| STAR | P2SYSTAR | 5,750 | 62.71 | 4,050 | 37.29 |
| THINK | P2SYTHIK | 1,950 | 21.91 | 7,850 | 78.09 |
| NEED | P2SYND | 3,700 | 42.33 | 6,150 | 57.67 |
| MILK | P2SYMLK | 8,250 | 84.59 | 1,600 | 15.41 |
| SCHOOL | P2SYSCHL | 5,150 | 56.67 | 4,700 | 43.33 |
| ALL GONE | P2SYALGN | 7,650 | 80.63 | 2,200 | 19.37 |
| IF | P2SYIF | 2,150 | 26.19 | 7,700 | 73.81 |
| HAT | P2SYHAT | 7,600 | 80.56 | 2,250 | 19.44 |
| PARTY | P2SYPRTY | 3,850 | 43.68 | 6,000 | 56.32 |

NOTE: The parent weight, W2R0, was used to obtain these statistics. Cell counts are unweighted to show the distribution in the ECLS-B 2-year data collection. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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

Table 10-7. Frequency distribution of parent reports of children’s syntax, 2-year data collection: 2003–04

| Variable name | Item description | Response option | Number | Weighted percent |
|---|---|---|---|---|
| P2CMBWRD | Combines words | (1) Not yet | 1,950 | 16.42 |
| | | (2) Sometimes | 3,500 | 34.11 |
| | | (3) Often | 4,400 | 49.47 |
| P2HOWCOM | How child communicates | (1) 1-word sentences | 1,900 | 21.55 |
| | | (2) 2-3 word phrases | 3,750 | 47.14 |
| | | (3) short sentences | 2,000 | 27.85 |
| | | (4) long sentences | 250 | 3.46 |
| P2PLURAL | Adds “s” to make plurals | (1) Yes | 4,700 | 64.05 |
| | | (2) No | 3,150 | 35.95 |
| P2TKOWSH | Adds “‘s” to talk about ownership | (1) Yes | 4,800 | 64.59 |
| | | (2) No | 3,050 | 35.41 |
| P2ADSING | Adds “ing” to verb to talk about activities | (1) Yes | 3,350 | 46.41 |
| | | (2) No | 4,500 | 53.59 |
| P2TKPST | Adds “ed” to talk about the past | (1) Yes | 2,150 | 30.56 |
| | | (2) No | 5,700 | 69.44 |

NOTE: The parent weight, W2R0, was used to obtain these statistics. Cell counts are unweighted to show the distribution in the ECLS-B 2-year data collection. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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

Table 10-8. Weighted means and standard deviations of children's total vocabulary scores for total sample and by key demographic variables, 2-year data collection: 2003–04

| | Average total vocabulary score | | |
|---|---|---|---|
| Characteristic | Number | Weighted mean | Standard deviation |
| Total sample | 9,850 | 20.06 | 11.87 |
| Child's race/ethnicity[1] | | | |
| White | 4,150 | 30.14 | 11.82 |
| Black | 1,500 | 28.66 | 11.33 |
| Hispanic, race specified | 1,350 | 27.78 | 11.46 |
| Hispanic, no race specified | 600 | 26.19 | 11.74 |
| Asian | 1,050 | 29.57 | 12.33 |
| Native Hawaiian/Pacific Islander | 50 | 29.44 | 14.03 |
| American Indian/Alaska Native | 250 | 26.73 | 11.16 |
| More than 1 race | 750 | 29.05 | 12.54 |
| Poverty status | | | |
| Below poverty threshold | 2,200 | 27.16 | 11.61 |
| At or above poverty threshold | 7,600 | 29.67 | 11.79 |
| Child's sex | | | |
| Male | 5,000 | 26.92 | 12.10 |
| Female | 4,800 | 31.44 | 11.00 |
| Child's age at assessment | | | |
| 21 months and under | # | 21.58 | 6.81 |
| 22–23 months | 950 | 26.03 | 11.25 |
| 24–25 months | 7,350 | 28.90 | 11.79 |
| 26–27 months | 1,100 | 32.01 | 11.48 |
| 28 months and over | 350 | 35.34 | 10.60 |
| Birth weight | | | |
| Normal | 7,200 | 29.44 | 11.74 |
| Moderately low | 1,500 | 26.29 | 11.67 |
| Very low | 1,050 | 19.84 | 11.37 |
| Mother's age (in years) | | | |
| 19 and under | 350 | 26.95 | 10.57 |
| 20–29 | 4,350 | 28.76 | 11.62 |
| 30–39 | 4,300 | 29.86 | 12.04 |
| 40 and over | 750 | 28.47 | 11.74 |

See notes at end of table.

Table 10-8. Weighted means and standard deviations of children's total vocabulary scores for total sample and by key demographic variables, 2-year data collection: 2003–04—Continued

| | Average total vocabulary score | | |
|---|---|---|---|
| Characteristic | Number | Weighted mean | Standard deviation |
| Mother's race/ethnicity[1] | | | |
| White | 4,550 | 30.08 | 11.80 |
| Black | 1,550 | 28.84 | 11.33 |
| Hispanic, race specified | 1,650 | 27.06 | 11.73 |
| Hispanic, no race specified | 50 | 26.83 | 12.19 |
| Asian | 1,200 | 29.20 | 12.34 |
| Native Hawaiian/Pacific Islander | 50 | 31.14 | 12.40 |
| American Indian/Alaska Native | 350 | 27.40 | 10.67 |
| More than 1 race | 250 | 28.35 | 11.88 |
| | | | |
| Mother's education | | | |
| 8th grade or below | 450 | 26.07 | 12.06 |
| 9–12th grades | 2,000 | 27.61 | 11.77 |
| High school diploma | 2,100 | 28.19 | 11.31 |
| Vocational/technical | 200 | 28.25 | 12.19 |
| Some college | 2,350 | 29.96 | 11.68 |
| Bachelor's degree | 1,600 | 30.26 | 11.97 |
| Graduate school (no degree) | 150 | 32.57 | 10.94 |
| Master's degree | 650 | 32.81 | 11.89 |
| Doctoral/professional degree | 250 | 33.94 | 10.02 |

# Rounds to zero.

[1] Race categories exclude Hispanic origin unless specified.

NOTE: The parent weight, W2R0, was used to obtain these statistics. Cell counts are unweighted to show the distribution in the ECLS-B 2-year data collection. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

Table 10-9. Weighted means and standard deviations for children’s language use/syntax by key demographic characteristics, 2-year data collection: 2003–04

| Demographic | Combines words (P2CMBWRD) | | | How communicates (P2HOWCOM) | | | Adds “s” to make plural (P2PLURAL) | | | Adds “s” for ownership (P2TKOWSH) | | | Adds “ing” to a verb for activities (P2ADSING) | | | Adds “ed” to refer to past (P2TKPST) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| characteristics | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD |
| Total sample | 9,850 | 2.33 | 0.74 | 7,850 | 2.13 | 0.78 | 7,850 | 1.36 | 0.48 | 7,850 | 1.35 | 0.48 | 7,850 | 1.54 | 0.50 | 7,850 | 1.69 | 0.46 |
| Mother’s race/ethnicity[1] | | | | | | | | | | | | | | | | | | |
| White | 4,600 | 2.46 | 0.70 | 3,900 | 2.16 | 0.80 | 3,900 | 1.32 | 0.47 | 3,900 | 1.31 | 0.46 | 3,900 | 1.52 | 0.50 | 3,900 | 1.69 | 0.46 |
| Black | 1,550 | 2.24 | 0.74 | 1,200 | 2.25 | 0.74 | 1,150 | 1.39 | 0.49 | 1,200 | 1.43 | 0.49 | 1,200 | 1.53 | 0.50 | 1,200 | 1.69 | 0.46 |
| Hispanic, race specified | 1,650 | 2.08 | 0.77 | 1,150 | 1.97 | 0.74 | 1,200 | 1.43 | 0.50 | 1,150 | 1.43 | 0.50 | 1,150 | 1.56 | 0.50 | 1,150 | 1.69 | 0.46 |
| Hispanic, no race specified | 50 | 2.08 | 0.76 | 50 | 2.05 | 0.55 | 50 | 1.14 | 0.35 | 50 | 1.37 | 0.47 | 50 | 1.52 | 0.50 | 50 | 1.69 | 0.46 |
| Asian | 1,200 | 2.23 | 0.77 | 1,000 | 2.16 | 0.79 | 1,000 | 1.57 | 0.50 | 1,000 | 1.43 | 0.50 | 1,000 | 1.60 | 0.49 | 1,000 | 1.75 | 0.43 |
| Native Hawaiian/ Pacific Islander | 50 | 2.29 | 0.71 | 50 | 2.17 | 0.70 | 50 | 1.35 | 0.48 | 50 | 1.31 | 0.46 | 50 | 1.49 | 0.50 | 50 | 1.77 | 0.42 |
| American Indian/ Alaska Native | 350 | 2.30 | 0.72 | 300 | 1.99 | 0.69 | 300 | 1.38 | 0.49 | 300 | 1.40 | 0.49 | 300 | 1.57 | 0.49 | 300 | 1.78 | 0.41 |
| More than 1 race | 250 | 2.37 | 0.73 | 200 | 2.17 | 0.82 | 200 | 1.41 | 0.49 | 200 | 1.41 | 0.49 | 200 | 1.56 | 0.50 | 200 | 1.72 | 0.45 |
| Poverty status | | | | | | | | | | | | | | | | | | |
| Below poverty threshold | 2,200 | 2.18 | 0.77 | 1,650 | 2.09 | 0.76 | 1,650 | 1.38 | 0.48 | 1,650 | 1.40 | 0.49 | 1,650 | 1.57 | 0.49 | 1,650 | 1.66 | 0.47 |
| At or above poverty threshold | 7,650 | 2.37 | 0.73 | 6,200 | 2.14 | 0.79 | 6,200 | 1.36 | 0.48 | 6,200 | 1.34 | 0.47 | 6,200 | 1.53 | 0.50 | 6,200 | 1.70 | 0.46 |

See notes at end of table.

Table 10-9. Weighted means and standard deviations for children's language use/syntax by key demographic characteristics, 2-year data collection: 2003–04—Continued

| Demographic | Combines words (P2CMBWRD) | | | How communicates (P2HOWCOM) | | | Adds "s" to make plural (P2PLURAL) | | | Adds "s" for ownership (P2TKOWSH) | | | Adds "ing" to a verb for activities (P2ADSING) | | | Adds "ed" to refer to past (P2TKPST) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| characteristics | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD |
| Child's race/ ethnicity[1] | | | | | | | | | | | | | | | | | | |
| White | 4,200 | 2.46 | 0.70 | 3,500 | 2.17 | 0.80 | 3,500 | 1.32 | 0.46 | 3,500 | 1.30 | 0.46 | 3,500 | 1.53 | 0.50 | 3,500 | 1.69 | 0.46 |
| Black | 1,550 | 2.23 | 0.74 | 1,150 | 2.23 | 0.73 | 1,150 | 1.39 | 0.49 | 1,150 | 1.43 | 0.50 | 1,150 | 1.54 | 0.50 | 1,150 | 1.69 | 0.46 |
| Hispanic, race specified | 1,350 | 2.14 | 0.77 | 1,000 | 1.98 | 0.73 | 1,000 | 1.40 | 0.49 | 1,000 | 1.41 | 0.49 | 1,000 | 1.56 | 0.50 | 1,000 | 1.68 | 0.47 |
| Hispanic, no race specified | 600 | 2.06 | 0.77 | 400 | 1.98 | 0.79 | 400 | 1.43 | 0.50 | 400 | 1.44 | 0.50 | 400 | 1.54 | 0.50 | 400 | 1.71 | 0.45 |
| Asian | 1,050 | 2.23 | 0.76 | 850 | 2.19 | 0.81 | 850 | 1.58 | 0.49 | 850 | 1.44 | 0.50 | 850 | 1.61 | 0.49 | 850 | 1.73 | 0.44 |
| Native Hawaiian/ Pacific Islander | 50 | 2.49 | 0.70 | 50 | 2.19 | 0.58 | 50 | 1.30 | 0.46 | 50 | 1.38 | 0.49 | 50 | 1.63 | 0.48 | 50 | 1.72 | 0.45 |
| American Indian/ Alaska Native | 250 | 2.30 | 0.73 | 200 | 1.92 | 0.74 | 200 | 1.37 | 0.48 | 200 | 1.44 | 0.50 | 200 | 1.63 | 0.48 | 200 | 1.74 | 0.44 |
| More than 1 race | 750 | 2.39 | 0.74 | 600 | 2.16 | 0.77 | 600 | 1.43 | 0.50 | 600 | 1.36 | 0.48 | 600 | 1.51 | 0.50 | 600 | 1.74 | 0.44 |
| Child's sex | | | | | | | | | | | | | | | | | | |
| Male | 5,050 | 2.22 | 0.76 | 3,800 | 2.00 | 0.78 | 3,800 | 1.39 | 0.49 | 3,800 | 1.40 | 0.49 | 3,800 | 1.61 | 0.49 | 3,800 | 1.74 | 0.44 |
| Female | 4,800 | 2.44 | 0.70 | 4,050 | 2.26 | 0.76 | 4,050 | 1.33 | 0.47 | 4,050 | 1.31 | 0.46 | 4,050 | 1.47 | 0.50 | 4,050 | 1.65 | 0.48 |
| Child's age at assessment | | | | | | | | | | | | | | | | | | |
| 21 months and under | # | 1.88 | 0.32 | # | 1.87 | 0.83 | # | 1.37 | 0.48 | # | 1.08 | 0.27 | # | 1.72 | 0.45 | # | 1.97 | 0.18 |
| 22–23 months | 950 | 2.25 | 0.76 | 750 | 1.97 | 0.75 | 750 | 1.42 | 0.49 | 750 | 1.43 | 0.49 | 750 | 1.61 | 0.49 | 750 | 1.76 | 0.43 |
| 24–25 months | 7,400 | 2.32 | 0.74 | 5,900 | 2.11 | 0.77 | 5,900 | 1.36 | 0.48 | 5,900 | 1.35 | 0.48 | 5,900 | 1.54 | 0.50 | 5,900 | 1.71 | 0.46 |
| 26–27 months | 1,100 | 2.41 | 0.71 | 950 | 2.29 | 0.82 | 900 | 1.33 | 0.47 | 900 | 1.32 | 0.47 | 900 | 1.51 | 0.50 | 900 | 1.61 | 0.49 |
| 28 months and over | 350 | 2.55 | 0.67 | 300 | 2.62 | 0.77 | 300 | 1.26 | 0.44 | 300 | 1.23 | 0.42 | 300 | 1.38 | 0.49 | 300 | 1.53 | 0.50 |

See notes at end of table.

Table 10-9. Weighted means and standard deviations for children’s language use/syntax by key demographic characteristics, 2-year data collection: 2003–04—Continued

| Demographic | Combines words (P2CMBWRD) | | | How communicates (P2HOWCOM) | | | Adds “s” to make plural (P2PLURAL) | | | Adds “s” for ownership (P2TKOWSH) | | | Adds “ing” to a verb for activities (P2ADSING) | | | Adds “ed” to refer to past (P2TKPST) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| characteristics | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD | Number | Mean | SD |
| Child’s birth weight | | | | | | | | | | | | | | | | | | |
| Normal | 7,200 | 2.35 | 0.74 | 6,000 | 2.15 | 0.79 | 6,000 | 1.36 | 0.48 | 6,000 | 1.35 | 0.48 | 6,000 | 1.53 | 0.50 | 6,000 | 1.69 | 0.46 |
| Moderately low | 1,500 | 2.18 | 0.76 | 1,200 | 1.98 | 0.74 | 1,200 | 1.39 | 0.49 | 1,200 | 1.44 | 0.50 | 1,200 | 1.62 | 0.49 | 1,200 | 1.75 | 0.43 |
| Very low | 1,050 | 1.81 | 0.78 | 650 | 1.71 | 0.70 | 650 | 1.54 | 0.50 | 650 | 1.52 | 0.50 | 650 | 1.70 | 0.46 | 650 | 1.86 | 0.34 |
| Mother’s age (in years) | | | | | | | | | | | | | | | | | | |
| 19 and under | 350 | 2.28 | 0.68 | 300 | 2.06 | 0.79 | 300 | 1.32 | 0.47 | 250 | 1.40 | 0.49 | 300 | 1.64 | 0.48 | 300 | 1.69 | 0.46 |
| 20–29 | 4,400 | 2.32 | 0.73 | 3,550 | 2.13 | 0.77 | 3,550 | 1.35 | 0.48 | 3,550 | 1.36 | 0.48 | 3,550 | 1.56 | 0.50 | 3,550 | 1.68 | 0.47 |
| 30–39 | 4,300 | 2.36 | 0.75 | 3,450 | 2.15 | 0.80 | 3,400 | 1.36 | 0.48 | 3,400 | 1.33 | 0.47 | 3,450 | 1.50 | 0.50 | 3,400 | 1.70 | 0.46 |
| 40 and over | 750 | 2.27 | 0.76 | 550 | 2.12 | 0.80 | 550 | 1.42 | 0.49 | 550 | 1.39 | 0.49 | 550 | 1.54 | 0.50 | 550 | 1.78 | 0.42 |
| Mother’s education | | | | | | | | | | | | | | | | | | |
| 8th grade or below | 450 | 1.93 | 0.75 | 300 | 1.91 | 0.78 | 300 | 1.47 | 0.50 | 300 | 1.46 | 0.50 | 300 | 1.57 | 0.50 | 300 | 1.63 | 0.48 |
| 9–12th grades | 2,000 | 2.24 | 0.74 | 1,550 | 2.06 | 0.75 | 1,550 | 1.36 | 0.48 | 1,550 | 1.39 | 0.49 | 1,550 | 1.57 | 0.50 | 1,550 | 1.65 | 0.48 |
| High school diploma | 2,100 | 2.29 | 0.73 | 1,650 | 2.12 | 0.77 | 1,650 | 1.35 | 0.48 | 1,650 | 1.38 | 0.48 | 1,650 | 1.56 | 0.50 | 1,650 | 1.71 | 0.45 |
| Vocational/technical | 200 | 2.23 | 0.77 | 150 | 2.08 | 0.78 | 150 | 1.34 | 0.47 | 150 | 1.37 | 0.48 | 150 | 1.62 | 0.49 | 150 | 1.70 | 0.46 |
| Some college | 2,350 | 2.39 | 0.73 | 1,900 | 2.15 | 0.77 | 1,900 | 1.36 | 0.48 | 1,900 | 1.33 | 0.47 | 1,900 | 1.53 | 0.50 | 1,900 | 1.73 | 0.45 |
| Bachelor’s degree | 1,600 | 2.43 | 0.74 | 1,300 | 2.18 | 0.80 | 1,300 | 1.37 | 0.48 | 1,300 | 1.32 | 0.47 | 1,300 | 1.51 | 0.50 | 1,300 | 1.72 | 0.45 |
| Graduate school (no degree) | 150 | 2.66 | 0.58 | 150 | 2.25 | 0.75 | 150 | 1.28 | 0.45 | 150 | 1.21 | 0.41 | 150 | 1.43 | 0.50 | 150 | 1.65 | 0.48 |
| Master’s degree | 700 | 2.57 | 0.66 | 600 | 2.33 | 0.87 | 600 | 1.32 | 0.47 | 600 | 1.30 | 0.46 | 600 | 1.42 | 0.49 | 600 | 1.66 | 0.47 |
| Doctoral/prof. degree | 250 | 2.62 | 0.60 | 200 | 2.27 | 0.83 | 200 | 1.39 | 0.49 | 200 | 1.31 | 0.46 | 200 | 1.41 | 0.49 | 200 | 1.63 | 0.48 |

# Rounds to zero.

[1] Race categories exclude Hispanic origin unless specified.

NOTE: The parent weight, W2R0, was used to obtain these statistics. Cell counts are unweighted to show the distribution in the ECLS-B 2-year data collection. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.

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.

# REFERENCES

Ainsworth, M.D.S., Blehar, M.C., Waters, E., and Wall, S. (1978). *Patterns of Attachment: A Psychological Study of the Strange Situation.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Aksan, N., and Kochanska, G. (2004). Links Between Systems of Inhibition from Infancy to Preschool Years. *Child Development, 75*: 1477–1499.

Andreassen, C., and Fletcher, P. (2005). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) Methodology Report for the Nine-Month Data Collection (2001–02), Volume 1: Psychometric Characteristics* (NCES 2005–100). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Ayres, A.J. (1979). *Sensory Integration and the Child.* Los Angeles: Western Psychological Services.

Baker, F. (2001). *The Basics of Item Response Theory,* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation (ERIC ED458219).

Bayley, N. (1993). *Bayley Scales of Infant Development, Second Edition*. San Antonio, TX: The Psychological Corporation.

Beals, D.E., and Snow, C.E. (1994). Thunder is When the Angels are Upstairs Bowling: Narratives and Explanations at the Dinner Table. *Journal of Narrative and Life History, 4:* 331–352.

Bethel, J., Green, J.L., Kalton, G., and Nord, C. (2005). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Methodology Report for the Nine-Month Data Collection (2001–02), Volume 2: Sampling* (NCES 2005–147). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Block, J. (1961). *The Q-Sort Method in Personality Assessment and Psychiatric Research.* Palo Alto, CA: Consulting Psychologists Press, Inc.

Block, J.H., and Block, J. (1980). The Role of Ego-Control and Ego-Resiliency in the Organization of Behavior. In W.A. Collins (Ed.), *The Minnesota Symposia on Child Psychology, 13*(pp. 39-101). Hillsdale, NJ: Lawrence Erlbaum Associates (Wiley).

Bock, R.D., and Zimowski, M.F. (1997). Multiple Group IRT. In W.J. van der Linden and R.K. Hambleton (Eds.), *Handbook of Item Response Theory* (pp. 433-448). New York: Springer-Verlag.

Bornstein, M.H., and Suess, P.E. (2000). Physiological Self-regulation and Information Processing in Infancy: Cardiac Vagal Tone and Habituation. *Child Development, 71*: 273–287.

Bowlby, J. (1969). *Attachment and Loss: Vol 1. Attachment.* New York: Basic Books.

Bowlby, J. (1973). *Attachment and Loss: Vol. 2 Separation: Anxiety and anger.* New York: Basic Books.

Bowlby, J. (1980). *Attachment and Loss: Vol. 3 Loss: Sadness and depression.* New York: Basic Books.

Brady-Smith, C., O'Brien, C., Berlin, L., and Ware, A. (1999). *Early Head Start Research and Evaluation Project 24-Month Child-Parent Interaction Rating Scales for the 3-Bag Assessment*. Unpublished manuscript. New York: Center for Children and Families, Teachers College, Columbia University.

Byrk, A.S., and Raudenbush, S.W. (1987). Application of Hierarchical Linear Models to Assessing Change. *Psychological Bulletin,* 101: 147–158.

Caldwell, B., and Bradley R.H. (1979). *Home Observation for Measurement of the Environment.* Little Rock, AR: University of Arkansas.

Caldwell, B., and Bradley R.H. (2001). *Home Inventory Administration Manual* (3rd Ed.). Little Rock, AR: University of Arkansas.

Cassidy, J., and Shaver, P.R. (Eds.). (1999). *Handbook of Attachment: Theory, Research, and Clinical Applications.* New York: The Guilford Press.

DeGangi, G.A., Poisson, S., Sickel, R.Z., and Wiener, A.S. (1995). *Infant/Toddler Symptom Checklist.* San Antonio, TX: The Psychological Corporation.

Feeney, J. (1999). Adult Romantic Attachment and Couple Relationships. In J. Cassidy and P.R. Shaver (Eds.), *Handbook of Attachment: Theory, Research, and Clinical Applications* (pp. 355–377). New York: The Guilford Press.

Fenson, L., Dale, P.S., Reznick, J.S., Bates, E., Thal, D., and Pethick, S. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development, 59*(5, Serial No. 242).

Fisher, A.G., Murray, E.A., and Bundy, A.C. (Eds.). (1991). *Sensory Integration: Theory and Practice.* Philadelphia, PA: F.A. Davis.

Gesell, A. (1949). *Gesell Developmental Schedules*. New York: The Psychological Corporation.

Green, P.J., Hoogstra, L.A., Ingels, S.J., Greene, H.N., and Marnell, P.K. (1997). *Formulating a Design for the ECLS: Review of Longitudinal Studies* (NCES 97–24). U.S. Department of Education, Washington, DC: National Center for Education Statistics Working Paper.

Greenberg, M.T., DeKlyen, M., Speltz, M.L., and Endriga, M.C. (1997). The Role of Attachment Processes in Externalizing Psychopathology in Young Children. *Attachment and Psychopathology*. (pp. 196-222). New York: Guilford Press.

Grossmann, K.E., Grossmann, K., and Zimmerman, P. (1999). A Wider View of Attachment and Exploration: Stability and Change During the Years of Immaturity (pp. 760-786). In J. Cassidy and P.R. Shaver (Eds.), *Handbook of Attachment: Theory, Research, and Clinical Applications*. New York: Guilford Press.

Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review, 9*: 139–150.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory* (p.42). Newbury Park, CA: Sage Publications, Inc.

Hart, B., and Risley, T.R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Paul H. Brookes Publishing Co.

Hazan, C., and Zeifman, D. (1999). PairBonds as Attachments: Evaluating the Evidence. In J. Cassidy and P.R. Shaver (Eds.), *Handbook of Attachment: Theory, Research, and Clinical Applications* (pp. 336–354). New York: The Guilford Press.

Ireton, H. (1997). *Child Development Inventory Manual.* Minneapolis, MN: Behavioral Science Systems.

Kochanska, G., Coy, K.C., and Murray, K.T. (2001). The Development of Self-regulation in the First Four Years of Life. *Child Development*, *72*: 1091–1111.

Kuder, G.F., and Richardson, M.W. (1937). The Theory of the Estimation of Test Reliability. *Psychometrika*, *2*:151–160.

Lamb, M.E. (2000). Attachment. In A.E. Kazdin (Ed.), *Encyclopedia of Psychology, 1*(pp. 284-289). Washington, DC: American Psychological Association.

Linacre, J.M., and Wright, B.D. (1994). Chi-Square Fit Statistics. *Rasch Measurement Transactions, 8*(2): 350.

Lord, F.M., and Novick, M. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Lyons-Ruth, K., and Jacobvitz, D. (1999). Attachment Disorganization. In J. Cassidy and P.R. Shaver (Eds.), *The Handbook of Attachment.* New York: Guilford Press.

Main, M. (2000). Attachment theory. In A.E. Kazdin (Ed.), *Encyclopedia of Psychology, 1*(pp. 289-293). Washington, DC: American Psychological Association.

Main, M., and Solomon, J. (1986). Discovery of an Insecure-disorganized/disoriented Attachment Pattern. In T.B. Brazelton, and M. Yogman (Eds.), *Affective Development in Infancy* (pp. 95–124). Norwood, NJ: Ablex.

Meisels, S.J., Atkins-Burnett, S., and Nicholson, J. (1996). *Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children* (NCES 96–18). U.S. Department of Education, Washington, DC: National Center for Education Statistics Working Paper.

Miller, A., McDonough, S.C., Rosenblum, K.L., and Sameroff, A.J. (2002). Emotion Regulation in Context: Situational Effects on Infant and Caregiver Behavior. *Infancy, 3*: 403–433.

Moore, K., Manlove, J., Richter, K., Halle, T., Le Menestrel, S., Zaslow, M., Greene, A.D., Mariner, C., Romano, A., and Bridges, L. (1999). *A Birth Cohort Study: Conceptual and Design Considerations and Rationale* (NCES 1999–01). U.S. Department of Education. Washington, DC: National Center for Education Statistics Working Paper.

Moss, E., St-Laurent, D., and Parent, S. (1999). Disorganized Attachment and Developmental Risk at School Age. In J. Solomon and C. George (Eds.), *Disorganized Attachment* (pp. 160–186). New York: Guilford.

Nelson, K. (1973). Structure and Strategy in Learning to Talk. *Monographs of the Society for Research in Child Development, 38*.

NICHD Early Child Care Research Network. (2004). Fathers' and Mothers' Parenting Behavior and Beliefs as Predictors of Children's Adjustment in the Transition to School. *Journal of Family Psychology, 18*: 628–638.

NICHD Early Child Care Research Network. (2005). Predicting Individual Differences in Attention, Memory and Planning in First Grade from Experiences at Home, Child Care and School. *Developmental Psychology, 41*: 99–114.

Nord, C., Edwards, B. Andreassen, C., Green, J.L., and Wallner-Allen, K. (2006). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), User's Manual for the ECLS-B Longitudinal 9-Month–2-Year Data File and Electronic Codebook* (NCES 2006–046). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Nord, C., Edwards, B., Hilpert, R., Branden, L., Andreassen, C. Elmore, A., Sesay, D., Fletcher, P. Green, J., Saunders, R., Dulaney R., Reaney, L., Flanagan, K., and West, J. (2004). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), User's Manual for the ECLS-B Nine-Month Restricted-Use Data File and Electronic Code Book* (NCES 2004–092). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Posada, G., Gao, Y., Wu, F., Posada, R., Tascon, J., Schoelmerich, A., Sagi, A., Kondo-Ikemura, K., Haaland, W., and Synnevaag, B. (1995). The Secure-base Phenomenon Across Cultures, Children's Behavior, Mothers' Preferences, and Experts' Concepts. In E. Waters, B.E. Vaughn, G. Posada, and K. Kondo-Ikemura (Eds.), *Caregiving, Cultural and Cognitive Perspectives on Secure-base Behavior and Working Models. Monographs of the Society for Research in Child Development, 60*(2-3, Serial No. 244).

Raju, N.S., van der Linden, W. J., and Fleer, P.F. (1995). IRT-Based Internal Measures of Differential Functioning of Items and Tests. *Applied Psychological Measurement*, *19*(4): 353–368.

Rathmann, P. (1994). *Good Night, Gorilla*. New York: Putnam

Raver, C.C. (2004). Placing Emotional Self-Regulation in Sociocultural and Socioeconomic Contexts. *Child Development, 75*: 346–353.

Smith, R.M., Schumacker, R.E., and Bush, M.J. (1998). Using Item Mean Squares to Evaluate Fit to the Rasch Model. *Journal of Outcome Measurement*, 2: 66-78.

Speltz, M.L., Greenberg, M.T., and DeKlyen, M. (1990). Attachment in Preschoolers with Disruptive Behavior: A Comparison of Clinic-referred and Nonproblem Children. *Development and Psychopathology*, *2*:31–46.

Stern, D. (1985). *The Interpersonal World of the Infant.* New York: Basic Books.

Stocking, M., and Lord, F.M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, 7: 201-210.

Trevarthen, C., and Aitken, K.J. (2001). Infant Intersubjectivity: Research, Theory and Clinical Applications. *Journal of Child Psychology and Psychiatry, 42*: 3–48.

van IJzendoorn, M.H., and Sagi, A. (1999). Cross-cultural Patterns of Attachment. In J. Cassidy and P.R. Shaver (Eds.), *Handbook of Attachment.* NY: Guilford Press.

von Davier, M., and von Davier, A.A. (2004). *A Unified Approach to IRT Scale Linking and Scale Transformations*. Princeton, NJ: Educational Testing Service.

Walker-Andrews, A.S. (1998). Emotions and Social Development: Infants' Recognition of Emotions in Others. *Pediatrics, 102* (5) Supplement: 1268–1271.

Ward, M.J., Lee, S.S., and Lipper, E.G. (2000). Failure-to-thrive Is Associated With Disorganized Infant-Mother Attachment and Unresolved Maternal Attachment. *Infant Mental Health Journal*, *21*:428–442.

Waters, E. (1995). The Attachment Q-Set (Version 3.0). In E. Waters, B.E. Vaughn, G. Posada, and K. Kondo-Ikemura (Eds.), *Caregiving, Cultural and Cognitive Perspectives on Secure-base Behavior and Working Models. Monographs of the Society for Research in Child Development*, *60*(2-3, Serial No. 244).

Waters, E., and Deane, K.E. (1985). Defining and Assessing Individual Differences in Attachment Relationships: Q-Methodology and the Organization of Behavior in Infancy and Early Childhood. In I. Bretherton and E. Waters (Eds.), *Growing Points in Attachment Theory and Research* (pp. 41–65), *Monographs of the Society for Research in Child Development*, *50*(1-2, Serial No. 209).

Waters, E., Vaughtn, B., Posada, G., and Kondo-Ikemura, K. (Eds.) (1995). Caregiving, Cultural, and Cognitive Perspectives on Secure-base Behavior and Working Models: New Growing Points of Attachment Theory and Research. *Monographs of the Society for Research and Child Development, 60*(2-3, Serial No. 244).

Willet, J.B. (1989). Questions and Answers in the Measurement of Change. In R.Z. Rothkopf (Ed.), *Review of Research in Education,* v.15 (pp. 350–351). Washington, DC: AERA.

Willet, J.B. (1997). Measuring Change: What Individual Growth Modeling Buys You. In E. Amsel and K.A. Renninger (Eds.), *Change and Development: Issues of Theory, Method, and Application* (p. 218). Mahwah, NJ: Lawrence Erlbaum Associates.

Zimowski, M.F., Muraki, E., Mislevy, R.J., and Bock, R.D. (1996). *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items.* Chicago: Scientific Software International.

*This page is intentionally left blank.*

# Appendix A: Intercorrelations of Major Direct Child Assessments

A-1. Intercorrelations of major direct child assessments, 2-year: 2003

| Item | X2MTLTSC | X2MTLSCL | X2MTL_F | X2MTL_G | X2MTL_H | X2MTL_I | X2MTL_J | X2MTRTSC | X2MTRSCL |
|---|---:|---:|---:|---:|---:|---:|---:|---:|---:|
| X2MTLTSC | † | .95 | .84 | .92 | .93 | .92 | .67 | .43 | .39 |
| X2MTLSCL | .95 | † | .88 | .96 | .98 | .97 | .71 | .36 | .46 |
| X2MTL_F | .84 | .88 | † | .93 | .82 | .76 | .34 | .36 | .44 |
| X2MTL_G | .92 | .96 | .93 | † | .96 | .92 | .52 | .35 | .43 |
| X2MTL_H | .93 | .98 | .82 | .96 | † | .99 | .70 | .34 | .42 |
| X2MTL_I | .92 | .97 | .76 | .92 | .99 | † | .77 | .32 | .41 |
| X2MTL_J | .67 | .71 | .34 | .52 | .70 | .77 | † | .22 | .31 |
| X2MTRTSC | .43 | .36 | .36 | .35 | .34 | .32 | .22 | † | .91 |
| X2MTRSCL | .39 | .46 | .44 | .43 | .42 | .41 | .31 | .91 | † |
| X2MTR_F | .36 | .40 | .46 | .40 | .36 | .34 | .18 | .80 | .86 |
| X2MTR_G | .38 | .44 | .44 | .43 | .41 | .40 | .27 | .91 | .99 |
| X2MTR_H | .38 | .45 | .43 | .43 | .42 | .41 | .30 | .91 | 1.00 |
| X2MTR_I | .36 | .43 | .38 | .40 | .41 | .41 | .32 | .86 | .96 |
| X2MTR_J | .33 | .41 | .33 | .36 | .39 | .40 | .37 | .78 | .89 |
| X2TBSPPT | .36 | .37 | .33 | .37 | .37 | .37 | .23 | .12 | .14 |
| C2SENSTV | .34 | .34 | .30 | .35 | .35 | .34 | .22 | .09 | .11 |
| C2POSRGD | .28 | .28 | .25 | .28 | .28 | .27 | .17 | .10 | .11 |
| C2NEGRGD | -.11 | -.12 | -.10 | -.12 | -.12 | -.12 | -.07 | -.01[1] | -.02 |
| C2NTRUSV | -.17 | -.17 | -.15 | -.17 | -.17 | -.17 | -.10 | -.06 | -.06 |
| C2COGDEV | .33 | .33 | .30 | .33 | .34 | .33 | .20 | .12 | .13 |
| C2DETACH | -.11 | -.11 | -.12 | -.12 | -.10 | -.10 | -.04 | -.06 | -.06 |
| C2ENGPRT | .44 | .46 | .43 | .46 | .46 | .44 | .27 | .17 | .21 |
| C2NEGPRT | -.22 | -.22 | -.21 | -.23 | -.22 | -.21 | -.13 | -.09 | -.10 |
| C2STNATT | .43 | .45 | .40 | .45 | .45 | .44 | .29 | .17 | .21 |
| X2CHHGHT | .01[1] | .09 | .08 | .08 | .08 | .08 | .06 | .02[1] | .12 |
| X2CHWGHT | -.05 | -.01[1] | .00[1] | -.02[1] | -.02 | -.03 | -.01[1] | .04 | .09 |
| X2CHMUAC | -.04 | -.02 | -.03 | -.04 | -.03 | -.03 | .02[1] | .07 | .09 |
| X2CHCRFM | .09 | .10 | .11 | .09 | .07[1] | .07[1] | .01[1] | .12 | .13 |
| X2CHBMI | -.06 | -.07 | -.05 | -.07 | -.08 | -.08 | -.05 | .04 | .02[1] |
| X2TASCLS | -.23 | -.24 | -.27 | -.27 | -.25 | -.23 | -.12 | -.18 | -.19 |
| X2TASCNF | .16 | .15 | .14 | .17 | .18 | .17 | .11 | .11 | .11 |
| X2TASSEC | .34 | .34 | .35 | .36 | .34 | .32 | .17 | .23 | .23 |
| X2TASDEP | -.15 | -.16 | -.14 | -.16 | -.16 | -.16 | -.12 | -.20 | -.21 |
| X2HSWARM | .24 | .24 | .25 | .25 | .24 | .22 | .11 | .11 | .11 |
| X2HSCOOP | .36 | .36 | .37 | .38 | .36 | .35 | .18 | .21 | .23 |
| X2HSENJY | .33 | .33 | .32 | .34 | .34 | .32 | .20 | .32 | .33 |
| X2HSINDP | .08 | .09 | .08 | .09 | .09 | .09 | .06 | .14 | .14 |
| X2HSATT | -.15 | -.16 | -.13 | -.15 | -.16 | -.16 | -.13 | -.17 | -.18 |
| X2HSUPST | -.19 | -.19 | -.17 | -.20 | -.20 | -.19 | -.13 | -.22 | -.23 |
| X2HSAVD | -.25 | -.25 | -.25 | -.26 | -.26 | -.24 | -.14 | -.21 | -.22 |
| X2HSDMND | -.31 | -.31 | -.30 | -.33 | -.32 | -.31 | -.19 | -.24 | -.25 |
| X2HSMDY | -.19 | -.18 | -.21 | -.20 | -.18 | -.17 | -.07 | -.11 | -.10 |

See notes at end of table.

A-1. Intercorrelations of major direct child assessments, 2-year: 2003—Continued

| Item | X2MTR_F | X2MTR_G | X2MTR_H | X2MTR_I | X2MTR_J | X2TBSPPT | C2SENSTV | C2POSRGD | C2NEGRGD |
|---|---|---|---|---|---|---|---|---|---|
| X2MTLTSC | .36 | .38 | .38 | .36 | .33 | .36 | .34 | .28 | -.11 |
| X2MTLSCL | .40 | .44 | .45 | .43 | .41 | .37 | .34 | .28 | -.12 |
| X2MTL_F | .46 | .44 | .43 | .38 | .33 | .33 | .30 | .25 | -.10 |
| X2MTL_G | .40 | .43 | .43 | .40 | .36 | .37 | .35 | .28 | -.12 |
| X2MTL_H | .36 | .41 | .42 | .41 | .39 | .37 | .35 | .28 | -.12 |
| X2MTL_I | .34 | .40 | .41 | .41 | .40 | .37 | .34 | .27 | -.12 |
| X2MTL_J | .18 | .27 | .30 | .32 | .37 | .23 | .22 | .17 | -.07 |
| X2MTRTSC | .80 | .91 | .91 | .86 | .78 | .12 | .09 | .10 | -.01[1] |
| X2MTRSCL | .86 | .99 | 1.00 | .96 | .89 | .14 | .11 | .11 | -.02 |
| X2MTR_F | † | .87 | .84 | .69 | .55 | .13 | .10 | .10 | -.02[1] |
| X2MTR_G | .87 | † | .99 | .94 | .83 | .14 | .11 | .11 | -.03 |
| X2MTR_H | .84 | .99 | † | .97 | .89 | .14 | .11 | .11 | -.02 |
| X2MTR_I | .69 | .94 | .97 | † | .96 | .13 | .10 | .10 | -.02[1] |
| X2MTR_J | .55 | .83 | .89 | .96 | † | .12 | .10 | .10 | -.02[1] |
| X2TBSPPT | .13 | .14 | .14 | .13 | .12 | † | .86 | .86 | -.28 |
| C2SENSTV | .10 | .11 | .11 | .10 | .10 | .86 | † | .63 | -.33 |
| C2POSRGD | .10 | .11 | .11 | .10 | .10 | .86 | .63 | † | -0.21 |
| C2NEGRGD | -.02[1] | -.03 | -.02 | -.02[1] | -.02[1] | -.28 | -.33 | -.21 | † |
| C2NTRUSV | -.04 | -.06 | -.06 | -.06 | -.06 | -.27 | -.33 | -.20 | .46 |
| C2COGDEV | .13 | .13 | .13 | .12 | .11 | .86 | .63 | .59 | -.18 |
| C2DETACH | -.07 | -.05 | -.05 | -.04 | -.04 | -.32 | -.31 | -.27 | .10 |
| C2ENGPRT | .19 | .20 | .21 | .19 | .18 | .66 | .60 | .53 | -.22 |
| C2NEGPRT | -.08 | -.10 | -.10 | -.09 | -.09 | -.27 | -.27 | -.23 | .35 |
| C2STNATT | .19 | .20 | .21 | .19 | .18 | .51 | .48 | .47 | -.17 |
| X2CHHGHT | .10 | .12 | .12 | .12 | .12 | .03 | .03 | .01[1] | -.01[1] |
| X2CHWGHT | .06 | .08 | .09 | .09 | .08 | -.03 | -.04 | -.02[1] | .01[1] |
| X2CHMUAC | .05 | .08 | .09 | .10 | .10 | -.08 | -.07 | -.06 | .01[1] |
| X2CHCRFM | .13 | .11 | .11 | .09 | .08 | .06[1] | .08[1] | .03[1] | -.01[1] |
| X2CHBMI | .01[1] | .02[1] | .02[1] | .02 | .02[1] | -.05 | -.06 | -.04 | .02[1] |
| X2TASCLS | -.20 | -.21 | -.21 | -.19 | -.17 | -.10 | -.08 | -.08 | .07 |
| X2TASCNF | .11 | .13 | .12 | .12 | .11 | .09 | .07 | .08 | -.01[1] |
| X2TASSEC | .24 | .24 | .24 | .22 | .19 | .19 | .18 | .15 | -.10 |
| X2TASDEP | -.19 | -.21 | -.21 | -.20 | -.18 | -.03 | -.03 | -.02 | -.02[1] |
| X2HSWARM | .12 | .12 | .11 | .10 | .09 | .16 | .15 | .12 | -.10 |
| X2HSCOOP | .22 | .23 | .23 | .21 | .18 | .19 | .18 | .14 | -.12 |
| X2HSENJY | .31 | .34 | .34 | .32 | .28 | .15 | .12 | .12 | -.04 |
| X2HSINDP | .14 | .15 | .15 | .13 | .12 | .01[1] | .01 | .01[1] | .02[1] |
| X2HSATT | -.15 | -.18 | -.19 | -.18 | -.17 | -.02[1] | -.02 | -.01[1] | -.00[1] |
| X2HSUPST | -.21 | -.24 | -.23 | -.22 | -.19 | -.04 | -.03 | -.03 | -.02[1] |
| X2HSAVD | -.20 | -.23 | -.23 | -.22 | -.19 | -.14 | -.10 | -.10 | .05 |
| X2HSDMND | -.23 | -.26 | -.26 | -.24 | -.22 | -.14 | -.12 | -.11 | .07 |
| X2HSMDY | -.12 | -.11 | -.11 | -.09 | -.08 | -.13 | -.13 | -.10 | .08 |

See notes at end of table.

A-1. Intercorrelations of major direct child assessments, 2-year: 2003—Continued

| Item | C2NTRUSV | C2COGDEV | C2DETACH | C2ENGPRT | C2NEGPRT | C2STNATT | X2CHHGHT | X2CHWGHT | X2CHMUAC |
|---|---|---|---|---|---|---|---|---|---|
| X2MTLTSC | -.17 | .33 | -.11 | .44 | -.22 | .43 | .01[1] | -.05 | -.04 |
| X2MTLSCL | -.17 | .33 | -.11 | .46 | -.22 | .45 | .09 | -.01[1] | -.02 |
| X2MTL_F | -.15 | .30 | -.12 | .43 | -.21 | .40 | .08 | .00[1] | -.03 |
| X2MTL_G | -.17 | .33 | -.12 | .46 | -.23 | .45 | .08 | -.02[1] | -.04 |
| X2MTL_H | -.17 | .34 | -.10 | .46 | -.22 | .45 | .08 | -.02 | -.03 |
| X2MTL_I | -.17 | .33 | -.10 | .44 | -.21 | .44 | .08 | -.03 | -.03 |
| X2MTL_J | -.10 | .20 | -.04 | .27 | -.13 | .29 | .06 | -.01[1] | .02[1] |
| X2MTRTSC | -.06 | .12 | -.06 | .17 | -.09 | .17 | .02[1] | .04 | .07 |
| X2MTRSCL | -.06 | .13 | -.06 | .21 | -.10 | .21 | .12 | .09 | .09 |
| X2MTR_F | -.04 | .13 | -.07 | .19 | -.08 | .19 | .10 | .06 | .05 |
| X2MTR_G | -.06 | .13 | -.05 | .20 | -.10 | .20 | .12 | .08 | .08 |
| X2MTR_H | -.06 | .13 | -.05 | .21 | -.10 | .21 | .12 | .09 | .09 |
| X2MTR_I | -.06 | .12 | -.04 | .19 | -.09 | .19 | .12 | .09 | .10 |
| X2MTR_J | -.06 | .11 | -.04 | .18 | -.09 | .18 | .12 | .08 | .10 |
| X2TBSPPT | -.27 | .86 | -.32 | .66 | -.27 | .51 | .03 | -.03 | -.08 |
| C2SENSTV | -.33 | .63 | -.31 | .60 | -.27 | .48 | .03 | -.04 | -.07 |
| C2POSRGD | -.20 | .59 | -.27 | .53 | -.23 | .47 | .01 | -.02 | -.06 |
| C2NEGRGD | .46 | -.18 | .10 | -.22 | .35 | -.17 | -.01[1] | .01[1] | .01[1] |
| C2NTRUSV | † | -.19 | .03 | -.24 | .45 | -.17 | .00[1] | .02[1] | .03 |
| C2COGDEV | -.19 | † | -.25 | .58 | -.22 | .54 | .02 | -.02[1] | -.07 |
| C2DETACH | .03 | -.25 | † | -.21 | .02[1] | -.14 | -.01[1] | -.01[1] | .00[1] |
| C2ENGPRT | -.24 | .58 | -.21 | † | -.37 | .76 | .00[1] | -.03 | -.06 |
| C2NEGPRT | .45 | -.22 | .02[1] | -.37 | † | -.38 | .02[1] | .03 | .03 |
| C2STNATT | -.17 | .54 | -.14 | .76 | -.38 | † | .01[1] | -.02[1] | -.05 |
| X2CHHGHT | .00[1] | .02 | -.01[1] | .00[1] | .02[1] | .01[1] | † | .51 | .27 |
| X2CHWGHT | .02[1] | -.02[1] | -.01[1] | -.03 | .03 | -.02[1] | .51 | † | .50 |
| X2CHMUAC | .03 | -.07 | .00[1] | -.06 | .03 | -.05 | .27 | .50 | † |
| X2CHCRFM | -.10 | .05 | -.01[1] | .02[1] | -.04[1] | .07[1] | .33 | .36 | .23 |
| X2CHBMI | .03 | -.04 | -.00[1] | -.04 | .02 | -.03 | -.07 | .82 | .41 |
| X2TASCLS | .07 | -.10 | .01[1] | -.16 | .14 | -.14 | .02 | .06 | .02 |
| X2TASCNF | -.03 | .09 | -.01[1] | .13 | -.07 | .13 | .01[1] | -.04 | -.03 |
| X2TASSEC | -.11 | .16 | -.04 | .24 | -.19 | .22 | -.01[1] | -.07 | -.05 |
| X2TASDEP | -.00[1] | -.02[1] | -.00[1] | -.08 | .03 | -.08 | -.02 | -.00[1] | -.03 |
| X2HSWARM | -.11 | .15 | -.05 | .19 | -.18 | .16 | -.04 | -.11 | -.08 |
| X2HSCOOP | -.14 | .17 | -.03 | .24 | -.23 | .21 | -.02 | -.10 | -.06 |
| X2HSENJY | -.04 | .14 | -.03 | .22 | -.12 | .22 | .00[1] | -.02[1] | .00[1] |
| X2HSINDP | .01[1] | .00[1] | -.00[1] | .04 | .02[1] | .05 | .04 | .03 | .04 |
| X2HSATT | .01[1] | -.02[1] | -.01[1] | -.07 | .05 | -.07 | -.00[1] | .03 | -.02[1] |
| X2HSUPST | .00[1] | -.05 | .00[1] | -.09 | .06 | -.09 | -.01[1] | .02 | -.02[1] |
| X2HSAVD | .05 | -.15 | .02[1] | -.18 | .13 | -.17 | .03 | .05 | .02 |
| X2HSDMND | .09 | -.12 | .01[1] | -.20 | .19 | -.18 | .02[1] | .08 | .04 |
| X2HSMDY | .08 | -.10 | .04 | -.15 | .10 | -.14 | -.01[1] | .03 | .04 |

See notes at end of table.

A-1. Intercorrelations of major direct child assessments, 2-year: 2003—Continued

| Item | X2CHCRFM | X2CHBMI | X2TASCLS | X2TASCNF | X2TASSEC | X2TASDEP | X2HSWARM | X2HSCOOP |
|---|---|---|---|---|---|---|---|---|
| X2MTLTSC | .09 | -.06 | -.23 | .16 | .34 | -.15 | .24 | .36 |
| X2MTLSCL | .10 | -.07 | -.24 | .15 | .34 | -.16 | .24 | .36 |
| X2MTL_F | .11 | -.05 | -.27 | .14 | .35 | -.14 | .25 | .37 |
| X2MTL_G | .09 | -.07 | -.27 | .17 | .36 | -.16 | .25 | .38 |
| X2MTL_H | .07[1] | -.08 | -.25 | .18 | .34 | -.16 | .24 | .36 |
| X2MTL_I | .07[1] | -.08 | -.23 | .17 | .32 | -.16 | .22 | .35 |
| X2MTL_J | .01[1] | -.05 | -.12 | .11 | .17 | -.12 | .11 | .18 |
| X2MTRTSC | .12 | .04 | -.18 | .11 | .23 | -.20 | .11 | .21 |
| X2MTRSCL | .13 | .02 | -.19 | .11 | .23 | -.21 | .11 | .23 |
| X2MTR_F | .13 | .01[1] | -.20 | .11 | .24 | -.19 | .12 | .22 |
| X2MTR_G | .11 | .02[1] | -.21 | .13 | .24 | -.21 | .12 | .23 |
| X2MTR_H | .11 | .02[1] | -.21 | .12 | .24 | -.21 | .11 | .23 |
| X2MTR_I | .09 | .02 | -.19 | .12 | .22 | -.20 | .10 | .21 |
| X2MTR_J | .08 | .02[1] | -.17 | .11 | .19 | -.18 | .09 | .18 |
| X2TBSPPT | .06[1] | -.05 | -.10 | .09 | .19 | -.03 | .16 | .19 |
| C2SENSTV | .08[1] | -.06 | -.08 | .07 | .18 | -.03 | .15 | .18 |
| C2POSRGD | .03 | -.04 | -.08 | .08 | .15 | -.02 | .12 | .14 |
| C2NEGRGD | -.01[1] | .02[1] | .07 | -.01[1] | -.10 | -.02[1] | -.10 | -.12 |
| C2NTRUSV | -.10 | .03 | .07 | -.03 | -.11 | -.00[1] | -.11 | -.14 |
| C2COGDEV | .05 | -.04 | -.10 | .09 | .16 | -.02[1] | .15 | .17 |
| C2DETACH | -.01[1] | -.00[1] | .01[1] | -.01[1] | -.04 | -.00[1] | -.05 | -.03 |
| C2ENGPRT | .02[1] | -.04 | -.16 | .13 | .24 | -.08 | .19 | .24 |
| C2NEGPRT | -.04[1] | .02 | .14 | -.07 | -.19 | .03 | -.18 | -.23 |
| C2STNATT | .07[1] | -.03 | -.14 | .13 | .22 | -.08 | .16 | .21 |
| X2CHHGHT | .33 | -.07 | .02 | .01[1] | -.01[1] | -.02 | -.04 | -.02 |
| X2CHWGHT | .36 | .82 | .06 | -.04 | -.07 | -.00[1] | -.11 | -.10 |
| X2CHMUAC | .23 | .41 | .02 | -.03 | -.05 | -.03 | -.08 | -.06 |
| X2CHCRFM | † | .20 | .11 | -.06[1] | -.08 | .04[1] | -.10 | -.06[1] |
| X2CHBMI | .20 | † | .05 | -.05 | -.07 | .01[1] | -.09 | -.09 |
| X2TASCLS | .11 | .05 | † | -.25 | -.58 | .39 | -.36 | -.50 |
| X2TASCNF | -.06[1] | -.05 | -.25 | † | .35 | -.22 | .19 | .26 |
| X2TASSEC | -.08 | -.07 | -.58 | .35 | † | -.30 | .74 | .83 |
| X2TASDEP | .04[1] | .01[1] | .39 | -.22 | -.30 | † | .21 | -.16 |
| X2HSWARM | -.10 | -.09 | -.36 | .19 | .74 | .21 | † | .73 |
| X2HSCOOP | -.06[1] | -.09 | -.50 | .26 | .83 | -.16 | .73 | † |
| X2HSENJY | -.12 | -.02[1] | -.60 | .38 | .68 | -.65 | .27 | .41 |
| X2HSINDP | .02[1] | .01[1] | -.31 | .19 | .32 | -.89 | -.26 | .02 |
| X2HSATT | .04[1] | .03 | .33 | -.17 | -.16 | .86 | .16 | -.25 |
| X2HSUPST | .05[1] | .03 | .36 | -.17 | -.29 | .90 | .05 | -.27 |
| X2HSAVD | .10 | .03 | .29 | -.14 | -.20 | -.19 | -.43 | -.31 |
| X2HSDMND | .13 | .08 | .58 | -.32 | -.60 | .66 | -.41 | -.65 |
| X2HSMDY | .01[1] | .03 | .36 | -.26 | -.86 | .06 | -.58 | -.56 |

See notes at end of table.

A-1. Intercorrelations of major direct child assessments, 2-year: 2003—Continued

| Item | X2HSENJY | X2HSINDP | X2HSATT | X2HSUPST | X2HSAVD | X2HSDMND | X2HSMDY |
|---|---:|---:|---:|---:|---:|---:|---:|
| X2MTLTSC | .33 | .08 | -.15 | -.19 | -.25 | -.31 | -.19 |
| X2MTLSCL | .33 | .09 | -.16 | -.19 | -.25 | -.31 | -.18 |
| X2MTL_F | .32 | .08 | -.13 | -.17 | -.25 | -.30 | -.21 |
| X2MTL_G | .34 | .09 | -.15 | -.20 | -.26 | -.33 | -.20 |
| X2MTL_H | .34 | .09 | -.16 | -.20 | -.26 | -.32 | -.18 |
| X2MTL_I | .32 | .09 | -.16 | -.19 | -.24 | -.31 | -.17 |
| X2MTL_J | .20 | .06 | -.13 | -.13 | -.14 | -.19 | -.07 |
| X2MTRTSC | .32 | .14 | -.17 | -.22 | -.21 | -.24 | -.11 |
| X2MTRSCL | .33 | .14 | -.18 | -.23 | -.22 | -.25 | -.10 |
| X2MTR_F | .31 | .14 | -.15 | -.21 | -.20 | -.23 | -.12 |
| X2MTR_G | .34 | .15 | -.18 | -.24 | -.23 | -.26 | -.11 |
| X2MTR_H | .34 | .15 | -.19 | -.23 | -.23 | -.26 | -.11 |
| X2MTR_I | .32 | .13 | -.18 | -.22 | -.22 | -.24 | -.09 |
| X2MTR_J | .28 | .12 | -.17 | -.19 | -.19 | -.22 | -.08 |
| X2TBSPPT | .15 | .01[1] | -.02[1] | -.04 | -.14 | -.14 | -.13 |
| C2SENSTV | .12 | .01 | -.02 | -.03 | -.10 | -.12 | -.13 |
| C2POSRGD | .12 | .01 | -.01 | -.03 | -.10 | -.11 | -.10 |
| C2NEGRGD | -.04 | .02[1] | -.00[1] | -.02[1] | .05 | .07 | .08 |
| C2NTRUSV | -.04 | .01[1] | .01[1] | .00[1] | .05 | .09 | .08 |
| C2COGDEV | .14 | .00[1] | -.02[1] | -.05 | -.15 | -.12 | -.10 |
| C2DETACH | -.03 | -.00[1] | -.01[1] | .00[1] | .02[1] | .01[1] | .04 |
| C2ENGPRT | .22 | .04 | -.07 | -.09 | -.18 | -.20 | -.15 |
| C2NEGPRT | -.12 | .02[1] | .05 | .06 | .13 | .19 | .10 |
| C2STNATT | .22 | .05 | -.07 | -.09 | -.17 | -.18 | -.14 |
| X2CHHGHT | .00[1] | .04 | -.00[1] | -.01[1] | .03 | .02[1] | -.01[1] |
| X2CHWGHT | -.02[1] | .03 | .03 | .02 | .05 | .08 | .03 |
| X2CHMUAC | .00[1] | .04 | -.02[1] | -.02[1] | .02 | .04 | .04 |
| X2CHCRFM | -.12 | .02[1] | .04[1] | .05[1] | .10 | .13 | .01[1] |
| X2CHBMI | -.02[1] | .01[1] | .03 | .03 | .03 | .08 | .03 |
| X2TASCLS | -.60 | -.31 | .33 | .36 | .29 | .58 | .36 |
| X2TASCNF | .38 | .19 | -.17 | -.17 | -.14 | -.32 | -.26 |
| X2TASSEC | .68 | .32 | -.16 | -.29 | -.20 | -.60 | -.86 |
| X2TASDEP | -.65 | -.89 | .86 | .90 | -.19 | .66 | .06 |
| X2HSWARM | .27 | -.26 | .16 | .05 | -.43 | -.41 | -.58 |
| X2HSCOOP | .41 | .02 | -.25 | -.27 | -.31 | -.65 | -.56 |
| X2HSENJY | † | .55 | -.45 | -.60 | -.41 | -.63 | -.43 |
| X2HSINDP | .55 | † | -.60 | -.67 | .36 | -.42 | -.28 |
| X2HSATT | -.45 | -.60 | † | .84 | -.07 | .69 | -.20 |
| X2HSUPST | -.60 | -.67 | .84 | † | .01[1] | .69 | -.06 |
| X2HSAVD | -.41 | .36 | -.07 | .01[1] | † | .24 | -.09 |
| X2HSDMND | -.63 | -.42 | .69 | .69 | .24 | † | .17 |
| X2HSMDY | -.43 | -.28 | -.20 | -.06 | -.09 | .17 | † |

† Not applicable.

[1] All correlations were significant at *p* < .05 except for these items.

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

## Appendix B: Toddler Attachment Sort Items

Below is a list of the TAS-45 items. Most of these items were derived from the two versions of the Attachment Q-sort (Waters & Deane, 1985; Waters, 1995). The wording of some of these items may have been modified from the original in order to make the item more readable. The Flesch-Kinkaid readability level of the items is generally at the eighth grade or below. In addition, a set of 6 items that address children's disorganized behaviors were culled from the literature on disorganized attachment and were added in order to be able to classify the disorganized attachment type.

The TAS-45 items include:

**Hot spot 1: Warm, Cuddly**

- Hugs and cuddles against mother without being asked to do so.
- Relaxes when in contact with mother.
- Seeks and enjoys being hugged by mother.
- When crying or upset, is easily comforted by contact with mother.

**Hot spot 2: Cooperativeness**

- When mother asks child to do something, child understands what she wants (may or may not obey).
- Cooperates with mother and gives her things if asked.
- Responds to positive hints from mother.
- Obeys when asked to bring or give mother something.
- When mother says "come here," child obeys.

**Hot spot 3: Enjoys Company**

- If asked, lets friendly adult strangers/new visitors hold or share toys.
- A social child who enjoys the company of others.
- Enjoys being hugged or held by friendly adult strangers/new visitors.
- Eager to join in with friendly adult strangers/new visitors, does not wait to be asked.
- Enjoys copying what friendly adult strangers do.

**Hot spot 4: Independent**

- Is very independent.
- Shows no fear, into everything.
- Usually finds something else to do when finished with an activity (does not go to mother for help.
- Takes off and explores new things on own.
- Hardly ever asks mother for any help (as child knows she is usually busy).

**Hot spot 5: Attention-seeker**

- Tries to stop mother from giving affection to other people (including family members).
- When mother talks with anybody else, child wants mother's attention.
- Wants to be at the center of mother's attention.
- When child is bored, will go to mother looking for something to do.
- Often wants mother's attention.

**Hot spot 6: Upset by separation**

- Is very clingy, stays closer to mother or returns more often than simply keeping track of mother's whereabouts.
- Gets upset if mother leaves and shifts to another place.
- Cries often, regardless of how hard or how long.
- Child does not try new things and always wants mother to help.
- Cries or tries to stop mother from leaving or moving to another place.

**Hot spot 7: Avoids others, does not socialize**

- Soon loses interest in friendly adult strangers/new visitors.
- Often plays out of mother's sight (e.g., watches TV-not needing mother).
- Turns away from friendly adult strangers/new visitors and goes own way.
- If there is a choice, child prefers to play with toys rather than with friendly adults.
- When a new visitor arrives, child first ignores or avoids him/her.

**Hot spot 8: Demanding/Angry**

- When child cries, cries loud and long.
- When child sees something really nice to play with, child will fuss and whine or try to drag mother over to it.
- When mother does not do what child wants right away, child knows she won't be "coming" and then fusses, gets angry or gives up.
- Easily becomes angry at mother.
- Cries as a way of getting mother to do what is wanted.

**Hot spot 9: "Moody"/Unsure about how to react/"unusual behaviors"**

- Generally cranky or grouchy when with mother.
- With mother, child suddenly switches mood. For instance goes from being nice to mean, or calm to upset (crying, afraid, angry), or gets upset and then goes blank.
- Goes all floppy (limp) when held by mother.
- Looks dazed and unsure (e.g., stares blankly, or freezes in an unusual position for a few seconds).
- Come to mother to give her toys, but will not touch or look at her.
- Suddenly aggressive towards mother for no reason (e.g., hits, slaps, pushes or bites mother).