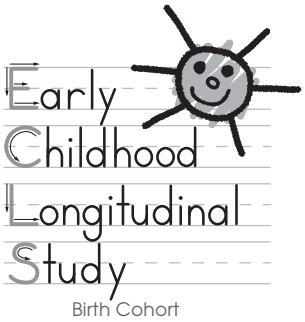
National Center for Education Statistics



U.S. Department of Education
Institute of Education Sciences
NCES 2005-100

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

Methodology Report for the Nine-Month Data Collection (2001–02)

Volume 1: Psychometric Characteristics

June 2005

Carol Andreassen
Philip Fletcher
Westat

Jerry West
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
Grover J. Whitehurst
Acting 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.

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

June 2005

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

Suggested Citation

Andreassen, C. and Fletcher, P. (2005). *Early Childhood Longitudinal Study, Birth Cohort (ECLS–B), Psychometric Characteristics*. Volume 1 of the *ECLS-B Methodology Report for the Nine-Month Data Collection* (NCES 2005–100). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

FOREWORD

This report describes the psychometric characteristics and related methodology of the 9-month data collection of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), which 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..

In the base year collection of the ECLS-B, when the children were about 9 months of age, the study interviewed parents (typically the biological mother), assessed children, and gathered information directly from the children's father figure. This report describes the psychometric instruments of the direct child assessments, including the Bayley Short Form–Research Edition (BSF-R), the Nursing Child Assessment Teaching Scale (NCATS), the physical measurements, and interviewer-completed observations of children's behavior. The report also describes some indirect assessments in the parent interview.

We hope that the information provided in this report will be useful to both researchers and policymakers. We further hope that the results reported here will encourage others to use the ECLS-B data, both now and in the future, as additional waves build upon this baseline.

Grover J. Whitehurst
Acting Commissioner
National Center for Education Statistics

This page is intentionally left blank.

ACKNOWLEDGMENTS

Over the past five 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 the 10,688 children and their parents who participated during the first wave of the study. 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 both during the national data collection and as a part of several earlier field tests of the study design and instrumentation. And, 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 were invaluable.

We wish to thank Natasha Cabrera of the University of Maryland at College Park (and formerly with the National Institute of Child Health and Human Development) for all her work in support of the health, special population, and father components of the study. We are especially grateful for her suggestions and encouragement in developing the father questionnaires of the ECLS-B, along with the assistance from the other members of the Developing a Daddy Survey (DADS) group.

Several others have consistently given their time in support of the study including: Vic Oliveira (Economic Research Service, U.S. Department of Agriculture); Michael Lopez, Louisa Tarullo, and Rachel Cohen (Administration on Children, Youth and Families, U.S. Department of Health and Human Services); Michael Kogan (Maternal and Child Health Bureau, U.S. Department of Health and Human Services); Linda Mellgren and Martha Moorehouse (Office of Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services); Karen Bourdon (National Institute of Mental Health, National Institutes of Health); and Howard Hoffman (National Institute on Deafness and Other Communication Disorders, National Institutes of Health).

Westat, under the direction of the National Center for Education Statistics (NCES), designed and conducted the 9-month wave of the study. NCES received day-to-day support from the Education Statistics Services Institute (ESSI), American Institutes for Research. NCES would like to recognize the contributions and efforts of several members of these two organizations' staffs. Carol Andreassen and Christine Nord of Westat led the instrument development activities. While the design of the ECLS-B child assessment battery and in particular the design of the Bayley Short Form–Research Edition were a team effort, no one invested more time, thought, and energy into these products than did Dr. Andreassen. Christine Nord contributed in a similar way to the parent interview and father questionnaires. NCES appreciates the work and dedication of each. They left a lasting imprint on the study that analysts will benefit from for many years to come. 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 for her tireless work with the NCHS and state Institutional Review Boards.

The design of the child assessment battery also benefited from the participation of expert workgroups. 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 in particular, and in Item Response Theory. These individuals included Drs. Don Rock and Judy Pollack of Educational Testing Service who provided Item Response Theory analysis expertise; Dr. Barbara Wasik of the Johns Hopkins University who advised on child assessment with low-socioeconomic status (SES) and language-minority children; and, Dr. Kathleen Matula who had been involved in the restandardization of the Bayley Scales of Infant Development, Second Edition for The Psychological Corporation and who advised on item administration and scoring and training of field staff. Dr. Kathryn Barnard of the University of Washington was consulted about the administration and coding of the Nursing Child Assessment Teaching Scale and generously provided a special coding training for Westat's trainers.

Kristin Denton Flanagan of ESSI and Liza Reaney (formerly of ESSI and now at Columbia University) supported NCES during the design and conduct of the ECLS-B and deserve special recognition. Each played an important role in the development of the ECLS-B and their efforts are reflected in the study instrumentation, data processing activities, and data products. NCES wishes to thank both of them for their years of service to the study.

Finally, a special thank you to Kendra Chandler Webb, age 9 (1994), for designing the ECLS logo.

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.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

List of Appendixes

List of Tables

TABLE OF CONTENTS (CONTINUED)

List of Tables (Continued)

TABLE OF CONTENTS (CONTINUED)

List of Tables (Continued)

List of Exhibits

TABLE OF CONTENTS (CONTINUED)

List of Figures

TABLE OF CONTENTS (CONTINUED)

List of Figures

TABLE OF CONTENTS (CONTINUED)

List of Figures (Continued)

This page is intentionally left blank.

1. INTRODUCTION

The Early Childhood Longitudinal Study project (ECLS) provides decisionmakers, researchers, child care providers, teachers, and parents with detailed information about children's early life experiences through two cohorts, the Birth Cohort (ECLS-B) and the Kindergarten Cohort (ECLS-K). The ECLS-B, the focus of this report, is a large scale nationally representative probability sample of children born in 2001 selected from a multistage sample design. The sample size for the base year data collection, which occurred when children were approximately 9 months of age, was 13,921. The overall response rate (10,688 parent completes) was 74.1 percent. 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 2005–113), available from NCES at http://nces.ed.gov/ecls.

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 for Children, Youth and Families (ACYF), and the U.S. Department of Agriculture (USDA).

1.1 Purpose of This Report

The ECLS-B is a rich data set that obtains information about a broad range of children's early experiences. Data collection is intended to continue through the end of first grade with the subsequent data collections at 2 years, 4 years, kindergarten, and the final data collection during first grade. This report documents the methodology for the child assessments in the ECLS-B, including the design, construction, implementation, quality control, and psychometric characteristics of the direct and indirect child assessment instruments in the 9-month data collection.

This chapter presents an overview of the purposes of child assessment in the ECLS studies in general and in the ECLS-B specifically, as well as a summary of the role of expert advisors in identifying appropriate instruments and in ensuring the quality of data collected from implementation in the field. A brief summary of the selected child assessments is provided, with an overview of the succession of field tests. For further information about the base year sample, the history and conceptual framework of the

other instruments in the ECLS-B, including the parent computer-assisted (CAPI) instrument, the variables in the 9-month public-use file, the overall design of the 9-month data collection, and the base year sample, see the following NCES documents available from http://nces.ed.gov/ecls: (1) *User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005–013) and (2) *ECLS-B Methodology Report for the Nine-Month Data Collection, Volume 2: Sampling* (NCES 2005–113).

In addition, two cross-sectional weights were used to obtain the data reported in this document. These two weights are W1R0 and W1C0. Weight W1R0 was used to estimate child-level characteristics associated with data collected through the parent interview and/or birth certificate. Examples relevant to child assessment include sets of questions in the parent interview addressing children's self-regulation behaviors, children's ages when developmental milestones were reached, parental knowledge of child development, and parents' child rearing beliefs. Weight W1C0 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 and/or birth certificate. 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.

1.2 Purpose of the Child Assessments in the Early Childhood Longitudinal Study

The ECLS-B is one study in the Early Childhood Longitudinal Study, which also includes the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) that follows children from kindergarten through fifth grade. A central goal of the ECLS studies, in general, 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. In order to understand children's growth and development, children need to be directly and indirectly assessed. Direct assessments provide invaluable objective information about the status and development of children. 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).

The purpose of assessment in the ECLS project, in general, is to provide descriptive data on children's current developmental status as well as on developmental growth. These data can be used to describe children's competencies and skills at different ages during the first 6 years of life and to provide valuable information on what most children in the United States are able to do in the domains of physical, cognitive and language, and social and emotional development at key points. Moreover, by assessing children at several points in time, it is possible to describe the levels and rates of growth for different groups of children. Information on children's development and experiences in the home, child care, early education programs, and schools is gathered through interviews with parents, child care providers, and teachers. Data from the child assessment battery complement these parent, child care provider, and teacher reports of children's experiences, providing important information about the children in the study such as their achievement relative to key developmental milestones (e.g., sitting up, first steps, first words).

The ECLS-B assessment is not intended to provide information on children's global mental ability or IQ. Instead, it is intended to provide descriptive information on those aspects of development that are deemed important for school readiness and early school success. The assessment is meant to provide descriptive data on children's early cognitive, motor, and social competencies—skills important for children as they begin their formal school careers.

In addition to providing rich descriptive information on children, the ECLS-B assessment enables researchers to accomplish several analytic goals. The data generated from the assessment can be used to explore the relationships between children's developmental outcomes and characteristics of their family, health care, child care, school, and community. The longitudinal nature of the study enables researchers to analyze children's physical, cognitive, social, and emotional growth and to relate trajectories of growth and change to variations in children's experiences. One main purpose of assessment in the ECLS, thus, is to provide data to permit direct examination of variation in children's cognitive skills, socioemotional status, and physical well-being by characteristics of children's in- and out-of-home experiences and background.

The design and implementation of the 9-month child assessments have been guided by consultation with child development and early education experts who participated in the Technical Review Panel and/or in specialized work groups that addressed specific technical needs. In addition, literature reviews (see section 1.4) were conducted to suggest the necessary content areas of the assessments and identify likely instruments. These activities are summarized in the following three

sections. An overview of the design of the child assessments and their field testing is presented in the final two sections of this chapter.

1.3 Technical Review Panel

The design of the content of the child assessment component of the ECLS-B has been guided by the Technical Review Panel, a panel of advisors comprised of representatives 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 Technical Review Panel 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 the design and implementation of 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 Technical Review Panel 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. Further details about the members of the ECLS-B Technical Review Panel for the 9-month data collection can be found in the *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Design and Operations Report for the Nine-Month Data Collection* (NCES, unpublished report).

Technical Review Panel 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 Technical Review Panel meetings. The members:

- 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 the quality of the implementation will ensure high data quality and valid results.

During the early design phases of the 9-month data collection, the Technical Review Panel meetings took place twice a year in Washington, DC. The Technical Review Panel 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 Technical Review Panel 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 Technical Review Panel work groups then reported back to the plenary group and their comments, suggestions, and recommendations were discussed and taken under review by NCES, the interagency partners, and Westat.

1.4 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); (2) *A Birth Cohort Study: Conceptual and Design Considerations and Rationale* (NCES Working Paper Series, Working Paper No 1999–01); and, (3) *Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children* (NCES Working Paper Series, Working Paper No. 96–18). Please refer to chapter 2 of the *User's Manual For The ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005–013) for further details. In addition, Child Trends reviewed available measures in 8 domains considered important for the ECLS-B, including the child assessments (NCES, unpublished manuscript). This review, in particular, was 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 samples to determine their operational feasibility for inclusion in the ECLS-B. At the same time, Westat staff also reviewed published direct assessments of children's developmental status (e.g., the Bayley Neurodevelopmental Screener, the Bayley Scales of Infant Development-II, the Mullen Scales of Early

Learning, the Denver Developmental Screening Test) to evaluate whether any of these were operationally feasible for administration by interviewers in a 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 in order 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. The review was also 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).

1.5 Overview of the Child Assessments in the Nine-Month Data Collection

As a result of the literature reviews and discussions with Technical Review Panel members, interagency partners, and NCES, it was decided that the direct assessment of the child was significant to the success of the 9-month data collection. Direct assessment would provide a firm baseline measure of overall developmental status against which data from later rounds could be compared. In particular, it was determined that it was important to include strong measures of infant developmental status, socioemotional functioning, and physical growth and well being.

In the developmental psychology literature, there has been disagreement about the continuity of mental development from infancy/toddlerhood through the early school years. On the one hand, some developmentalists claim that continuity of mental development cannot be established until the acquisition of language. On the other hand, some developmental researchers have shown continuity between key early emerging cognitive processes and later cognitive outcomes. However, the measurement of these early cognitive processes is usually laboratory based and requires specialized technical equipment, neither of which was feasible for the ECLS-B (see Sigman and Bornstein 1986, for a review). Therefore the

assessment of cognitive functioning in the ECLS-B required a reputable standardized test that was feasible for implementation in the field by administrators untrained in standardized assessment and that required no specialized technical equipment. The predictive ability of most standardized measures of cognitive functioning from infancy to the early school years is low to moderate for cognitive functioning, although predictive ability for psychomotor development tends to be greater. Consequently, it was important to select a standardized assessment that provided the best predictive ability possible. To measure developmental status, the consensus of NCES and the Technical Review Panel was to obtain both the mental and psychomotor scales of the BSID-II, which is described in chapter 2.

Advisors also recommended that children's socioemotional functioning be assessed with the Nursing Child Assessment Teaching Scale (NCATS), an observational measure of parent-child interaction that obtains separate scores for the parent and for the child on various aspects of interaction behaviors. 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) and other health studies conducted by the World Health Organization (WHO). These standard measurements include children's length, weight, middle upper arm circumference, and head circumference in the case of very low birth weight babies (i.e., 1,500 grams or less).

Direct assessments, administered by the interviewer, are effective means for obtaining a snapshot of the child's functioning at a single point in time, i.e., on the day of the home visit. For example, during the 9-month home visit, the interviewer observes and rates whether and how much the child cried during the direct assessment and therefore may not have done as well as possible during the direct assessments that day. The indirect measurements of children's functioning supplement the direct assessments by obtaining information about children's typical functioning.

There are two types of indirect measurements, which are obtained from the parent's point of view. The first type obtains information about children's typical functioning on an everyday basis. For example, the parent respondent is asked to indicate how frequently (on a 7-point scale ranging from "never" to "is like this most times") the child behaves in such ways as "startles easily," or "is irritable or fussy." The second type obtains information about children's environments and early experiences, such as parents' knowledge about child development, their parenting attitudes, and typical activities with the child. Such supplemental information could be useful in explaining children's current functioning during

the direct assessments as well as longitudinally as subsequent outcome measures become available. Taken together, the direct and the indirect measurements provide an in-depth view of children's typical functioning and typical environment.

The indirect child assessments in the ECLS-B were recommended to obtain these two types of information. Two indirect measurements were suggested as a continuation of the assessment of children's developmental functioning, although this time by indirect measurement obtained through parent report. These two indirect measurements include a set of questions about the age at which the child achieved key developmental milestones, and a set of questions that assess children's ability to self-regulate and self-soothe themselves. The second type of indirect measurement obtains information about the child's home experiences and home environment information. This information is obtained partly through interviewer observations and partly through several sets of questions to the parent respondent.

1.6 Field Testing

Two field tests of the entire 9-month protocol were conducted, as well as two pilot studies to investigate the feasibility of redesign of the parent interview and of the implementation of a shortened and streamlined BSID-II. The first field test of the entire home visit protocol was carried out in the fall of 1999. This field test demonstrated that the home visit was too burdensome to participants and to field representatives. As a result, the 9-month instruments were redesigned to create a simpler, more streamlined, and less burdensome home visit. Over the better part of the following year, the 9-month home visit was redesigned. Two rounds of brief, small scale pilot testing were conducted to assess the feasibility of administration of a shortened BSID-II in the home setting by field interviewers, to assess revisions to the field interviewer training, and to assess a streamlined parent interview. The second field test of the new redesigned home visit protocol was carried out in the fall of 2000 with results suggesting that the streamlined design was successful. Further details about outcomes of the field tests with respect to the direct assessments are presented in their respective sections about each assessment.

1.7 Organization of This Report

Subsequent chapters describe the specific instruments and sets of questions included in the 9-month ECLS-B data collection and provide an overview summary of 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 decision that led up to the development of the shortened version of that assessment, the BSF-R. Chapters 3 through 5 describe the work that was done to develop the BSF-R and summarizes the BSF-R variables in the public-use data file. 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 then summarizes the interviewer observations of children's behaviors and of their home environment. Chapters 9 and 10 summarize several sets of questions to parents about children's behaviors and their knowledge about child development and parenting beliefs. A list of references is provided at the end of the chapters, together with a table of intercorrelations of the direct child assessments (appendix A).

This page is intentionally left blank

2. DIRECT CHILD ASSESSMENTS

This chapter focuses on the instruments that directly assess children's status and performance. The 9-month data collection included three different direct assessments of children's developmental status, socioemotional functioning, and physical growth and development. Children's developmental status was assessed with a shortened form of the Bayley Scales of Infant Development, Second Edition (BSID-II), called the Bayley Short Form–Research Edition (BSF-R). The BSF-R mental scale includes items designed to assess children's cognitive functioning in such areas as vocalization and receptive language, object permanence, problem solving and exploration of objects, and so forth. The BSF-R motor scale includes items designed to assess fine motor skills, such as grasping and eye-hand coordination, and gross motor skills, such as dynamic movement and attainment of motor milestones. The Nursing Child Assessment Teaching Scale (NCATS) assesses children's socioemotional functioning within the parent-child interaction context, as well as the caregiver's sensitivity and responsivity to the child, and provision of cognitive growth fostering support. Standard measurements of physical growth are also included. The child's length is obtained with the child lying supine on a measurement mat. The child's weight is obtained while being held by the mother, with the child's weight being subtracted, or tared, from the mother's weight. The child's middle upper arm circumference is also obtained. This measurement is widely used by health studies sponsored by the World Health Organization (WHO), as a concise indicator of nutritional status. The child's upper arm length is obtained and the circumference of the upper arm is obtained at the midpoint. Finally, in the case of children born at very low birth weight (1,500 grams or less), head circumference is also obtained. Each of these assessments is described in detail in the following sections.

2.1 Bayley Scales of Infant Development, Second Edition

The ECLS-B originally intended to obtain the full BSID-II at the 9-month data collection and at subsequent data collections for which the BSID-II would be age-appropriate. However, the burden to interviewers and to participants that was found in the fall 1999 field test led to the design of a shortened and streamlined BSID-II called the Bayley Short Form–Research Edition (BSF-R). The BSF-R was implemented in the 9-month national data collection.

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

A key objective of the ECLS-B is to describe children's growth and development from infancy to the early school years. In order to describe children's developing skills, it was necessary to select a measure that provided a comprehensive snapshot of children's varying skills at multiple ages. In addition, because of the need for a strong anchoring data point, 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. 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 represent behaviors and responses geared to the identification of pathology rather than the full range of developmental abilities.

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 BSID-II, 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 so that it could be administered at the 9-month and at the 2-year data collection points to obtain continuity of measurement of growth in the ECLS-B.

Secondly, the BSID-II has the advantage of being among the more psychometrically sound standardized assessments available for infants and young children. Critical psychometric properties of any standardized test include the precision of scores, stability of scores over time, and predictive validity. These issues were especially critical in view of the wide range of infant development that must be assessed within ECLS-B's longitudinal framework.

Review of the psychometric properties presented in the BSID-II manual (Bayley, 1993) showed acceptable internal consistency (coefficient alpha), and acceptable test-retest reliability (with a small sample of 175 children). Evidence for concurrent validity was demonstrated by moderate to strong correlations with other standardized assessments, such as the McCarthy Scales of Children's Abilities (MSCA), and the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R). The BSID-II manual rightly asserts that predictive validity is a more complex issue requiring cumulative cross-study evidence and is not reducible to the numeric value of a single variable. That said, most evidence for

predictive validity has been modest and was obtained using the first edition of the Bayley Scales of Infant Development. In addition, the scales' predictive powers tend to increase with the child's age and for children scoring at the upper and lower ends of the ability distribution. However, this is consistent across developmental assessments.

Pertinent psychometric information about the reliability and validity of the BSID-II is summarized from the *Bayley Scales of Infant Development Manual* (Bayley 1993) and presented in table 1 (reliability) and table 2 (validity). Table 1 presents coefficient alpha across the entire age range of the BSID-II for the mental scale and for the motor scale, as well as the average coefficient alpha for each scale. These coefficients were obtained on the standardization data set in which children were grouped into 17 age groups with n = 100 for each age group. (Please see table 3 for a listing of the age groups in the standardization data set.) Coefficient alpha is a measure of the internal consistency of a scale. That is, coefficient alpha assesses how well a set of items measures a single construct. Alpha coefficient ranges in value from 0 to 1. A high value of alpha (closer to 1) provides evidence that the items are consistent and are measuring the same underlying construct. When the value of alpha is low (closer to 0), it indicates that the items have a multidimensional structure, coefficient alpha will usually be low (closer to 0). The values of coefficient alpha in table 1 demonstrate that the BSID-II is a highly reliable instrument.

Table 1 also presents the standard error of measurement (*SEm*) of the mental scale and of the motor scale. The *SEm* provides an estimate of the amount of error in an individual's observed test score and is inversely related to the reliability coefficient so that the greater the reliability, the lower the standard error. The *SEm* forms the basis of confidence intervals around the individual's obtained score. The values of *SEm* in table 1 for the mental and for the motor scale demonstrate that the *SEm* is acceptable.

Table 1 summarizes the BSID-II's short term test-retest reliability, or stability. The stability of scores was obtained on a subsample of 175 children from the standardization sample who were tested twice. Test-retest intervals were in the range of 1–16 days apart, with a median interval of 4 days. Stability coefficients were obtained for children at ages 1 month, 12, 24, and 36 months. Data were then pooled for ages 1 month and 12 months, and for ages 24 months and 36 months. Only data on ages 1 month and 12 months (n = 90) are presented in table 1 because they are closest to the target age of the ECLS-B 9-month data collection. These results show that at these younger ages and over a short time interval, the BSID-II has a suitable degree of reliability for the purpose of the current assessments.

Table 1. Reliability coefficients and standard errors for the Bayley Scales of Infant Development-II (BSID-II): 1993

	Coefficient alpha		Standard errors (SEm)[1]	
Scale	Range	Average r^2 (Fisher's *z* transformation)	Range	Average SEm
Mental scale	.78–.93	.88	3.90–7.04	5.21
Motor scale	.75–.91	.84	4.47–7.56	6.01

Test-retest stability coefficients

	Test		Retest		
	Mean	SD	Mean	SD	Correlation
Mental scale	100.21	14.77	103.26	16.72	.83
Motor scale	98.76	15.08	100.53	15.54	.77

[1]SEm in scaled score units

SOURCE: Bayley, N. (1993). *Bayley Scales of Infant Development Manual.* San Antonio, Texas: The Psychological Corporation.

Table 2 presents statistics related to the test validity of the BSID-II mental and motor scales with other standardized measures. The first part compares the BSID-II mental and motor subscales with the mental and motor scales of first edition of the BSID. In addition to strong correlations between the mental and motor scales of the BSID-II and the BSID, these values also demonstrate the psychometric phenomenon known as the "Flynn Effect," which refers to the steady rise in scores on standardized assessments of approximately 3 points per decade (assuming a mean of 100 and standard deviation of 15) (Flynn 1984). It was this increase in BSID scores over the decades that necessitated its renorming in 1993. Accordingly, BSID scores are higher than BSID-II scores, which were recently renormed.

The second part compares the BSID-II with the subscales of the McCarthy Scales of Children's Abilities (MSCA), the Wechsler Primary and Preschool Scales of Intelligence-Revised Edition (WPPSI-R), and with the Differential Abilities Scale (DAS). These statistics show that the BSID-II mental scale is moderately correlated with other well-established assessments of cognitive functioning, such as the MSCA and the WPPSI-R, although the correlations with the DAS were more modest. The correlations of the BSID-II motor scale with these assessments are also rather modest.

Table 2. Concurrent validity of the Bayley Scales of Infant Development, Second Edition (BSID-II) (1993) with the first edition of the Bayley Scales of Infant Development (BSID), the McCarthy Scales of Children's Abilities (MSCA), Wechsler Preschool and Primary Scales of Intelligence–Revised (WPPSI-R), and the Differential Abilities Scale (DAS): 1993

Validity of BSID and BSID-II (N = 200)	Mean	SD	Correlation (r)
Mental Scale			
BSID	111.6	17.2	.62
BSID-II	99.8	14.9	.62
Motor Scale			
BSID	110.5	15.3	.63
BSID II	100.4	16.2	.63

BSID-II and McCarthy Scales of Children's Abilities (MSCA) (N = 30)

	Correlation (r) with BSID-II	
MSCA Subscale	Mental Scale	Motor Scale
Verbal	.77	.45
Performance	.69	.44
Quantitative	.59	.33
Memory	.62	.18
Motor	.57	.59
General cognitive	.79	.45

BSID-II and Wechsler Preschool and Primary Scales of Intelligence–Revised (WPPSI-R) (N = 40)

	Correlation (r) with BSID-II	
WPPSI-R Scale	Mental Scale	Motor Scale
Performance IQ	.63	.37
Verbal IQ	.73	.39
Full scale IQ	.73	.41

BSID-II and Differential Ability Scales (DAS) (N = 25)

	Correlation (r) with BSID-II	
DAS Scale	Mental Scale	Motor Scale
General concept	.49	.35
Nonverbal comprehension	.30	.24

SOURCE: Bayley, N. (1993). *Bayley Scales of Infant Development Manual*. San Antonio, Texas: The Psychological Corporation.

One aim for the ECLS-B is to obtain data that can be compared with data from other nationally representative, large scale studies, including the National Longitudinal Survey of Youth (NLSY79), and the Children of the NLSY79, the Comprehensive Child Development Project, the Early Head Start Research and Evaluation Project, and the National Institute of Child Health and Human Development (NICHD) Study of Early Child Care. The BSID-II, accordingly, has the advantage of having been used in other federally sponsored studies of child development, such as the 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.

The utility of the BSID-II for the ECLS-B is complemented by the additional and different exposure that will be given it. For instance, previous studies have used a single BSID-II assessment to predict later child outcome results. The general finding has been that infant measures of ability are only modestly predictive of children's later intellectual growth. The ECLS-B, however, obtains scores at two separate data collections, 9 months and 2 years, which will enable analysts to examine developmental status as a repeated measure and analyze growth rates and how they predict later outcomes. Therefore, the ECLS-B can make an important contribution to understanding the impact of contextual influences on children's growth rates and their associations with later school outcomes.

2.1.2 Description of the BSID-II

It was originally intended that the full BSID-II be administered during the 9-month home visit. The focus of the BSID-II is on the measurement of ability at the earliest ages, as manifest in gross and fine motor abilities, perceptual-motor integration, perception, habituation, pre-verbal communication, learning, problem solving, language, reasoning, concept attainment, and so on. The BSID-II is an individually administered instrument that assesses the current developmental functioning of infants and children from 1 month to 42 months. In total, the BSID-II is composed of two scales, the mental scale and the motor scale. The entire mental scale consists of 178 items that assess abilities such as memory, habituation, problem solving, ability to vocalize, language, and social skills. The entire motor scale consists of 111 items that assess fine motor abilities (such as grasping and pre-writing skills) and gross motor abilities (such as rolling, crawling and creeping, sitting, standing, walking, running, and jumping). All 178 mental items and 111 motor items would never be administered to a child. Items are organized into developmentally appropriate age sets and only relevant age sets would be administered. The number of items in each age set varies. All the items in the BSID-II are arranged in the order of their

developmental difficulty. Most of the items are administered to the child, although there is a small percentage that can be scored by observation of the child during the administration of other items.

As described in the BSID-II manual (Bayley 1993), a panel of experts categorized the items in the mental scale into four "facets," or content areas, including cognition, language, interpersonal/social, and motor facets. There are no norms available for these facets, only age equivalents, which limits their usefulness. In addition, a factor study by Burns, Burns, and Kabacoff (1992) showed that up to about the 26–28-month age set, the mental scale of the BSID-II is predominantly unidimensional, the main factor being what could be characterized as "ability." However, beyond that age, due to the maturation of the brain and the divergence of skills that are tapped by a developmental assessment, several factors were identified, suggesting that item categories and breadth of content become more important at the later ages.

The BSID-II specifies sets of items to administer to a child depending on his or her chronological age. For example, the mental scale item set specified for a 9-month-old child includes 21 administered items with 4 items scored by observation, and 13 administered items on the motor scale with one item scored by observation. In most cases, administration of the age-appropriate item set is sufficient to obtain an accurate assessment of a child's abilities.[1] In some cases, however, it is necessary to administer additional sets of items in order 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 (and 2 or fewer "no credits" on the motor scale) within their item age 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 manual, administration of the age-appropriate BSID-II requires at least 30 minutes for children less than 15 months of age. For older children, the administration time can be at least an hour. Additional time is required, of course, if additional item age sets need to be administered to establish the basal or ceiling.

Raw scores obtained from the number of passed and failed mental ability items and motor ability items are then converted into a Mental Development Index (MDI) and a Psychomotor Development Index (PDI), for the mental scale and for the motor scale, respectively. Both the MDI and

[1] Because the bidirectionality of the basal and ceiling rules (i.e., 5 or fewer credits versus 3 or fewer no credits on the mental scale) can be confusing to apply, these rules were simplified for the BSF-R. These simplified BSF-R basal and ceiling rules are presented in section 2.1.7.

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, e.g., the Wechsler Preschool and Primary Scales of Intelligence. Conceptually, however, the BSID-II should be thought of as an assessment of developmental status rather than of intelligence because children this age have not developed the verbal skills typically included in most intelligence tests. 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 Behavior Rating Scale (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), 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 with the examiner, 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 may be due, at least in part, to poor control over motor responding or emotional regulation.

2.1.3 Initial Plan: Full BSID-II as Implemented in the National Early Head Start Research and Evaluation Project

As initially designed, it was the intention of NCES that the full BSID-II, including the mental scale, motor scale, and the Behavior Rating Scale, be administered to all sampled children in the ECLS-B at the earliest data collection points. For this purpose, Westat acquired from Mathematica Policy Research (MPR) its complete protocol and supporting materials for the BSID-II administration that was implemented in the national Early Head Start Evaluation and Research Project to serve as a model for the development of materials for the ECLS-B to the extent possible. The materials could not be adopted in their entirety because the national Early Head Start Research and Evaluation Project did not administer the motor scale. Furthermore, although the complete mental scale was administered to all the children in

the sample, MPR modified the administration by having all administrations begin at the same age set starting point. For example, at the 14-month data collection, all children in the 13 to 15 month age range began with the 11-month set of items. Depending upon performance, then, children then either completed the mental scale with that set of core items or proceeded to the basal item set or the ceiling item set. In the ECLS-B, however, the BSF-R had a core set of items based on the 9-month age set that was administered to all children. To the extent possible, adherence to the same procedures and materials used in the EHS Evaluation Study would help to ensure the comparability of data between the two studies, at least for the mental scale.

2.1.4 Fall 1999 Field Test Design

The first field test of the 9-month design began in September 1999 with a sample of approximately 1,500 cases. The complete home visit protocol was implemented in the Fall 1999 Field Test and included the computer-assisted parent interview, the parent self-administered questionnaire (of items deemed sensitive), the Letter-Word Recognition subtest of the Woodcock-Johnson for the parent, the full BSID-II for the child, the videotaped administration of the Nursing Child Assessment Teaching Scale for the parent and child (NCATS; see, Chapter 6, below, for details), the physical measurements, the resident father questionnaire (or the nonresident father package, as appropriate), and the short form of the HOME observations, as well as the Child Observations (consisting of the entire BRS from the BSID-II), which were completed by the interviewer after the home visit. It was intended that the ECLS-B home visit at 9 months would take about 90 to 120 minutes, with approximately 35 minutes of that time allotted to the direct child assessment components (the BSID-II, the NCATS, and the physical measurements).

Problems implementing the design of the home visit in the fall 1999 field test soon became apparent in a number of areas. First, interviewer attrition was at twice the expected rate. At the September 1999 training, 24 interviewers attended and 22 successfully completed training. Yet after 2 months, there were only 15 interviewers still working, for an attrition rate of about 32 percent, which compares with an expected annual attrition rate of only 15 percent. A second training was held in November, which 18 trainees passed. But after 6 weeks, only 13 of the 18 were still working on ECLS-B. Judging by their comments and feedback to Westat, interviewers seemed to feel overwhelmed by the sheer number of the tasks administered during the home visit and by their complexity. Simply put, interviewers felt they were being asked to do too much and that they had not been adequately prepared for the demands of the tasks.

The second indicator of problems in the design of the 9-month home visit was the low level of production in the field. It had been estimated that interviewers would work an average of 20 hours a week in the field, completing two cases a week for the 4½-month field test period. This estimate was based on a 60-minute parent interview plus a 30–35 minute direct child assessment that would include the parent-child videotaped interaction and physical measurements. Instead, the parent interview was taking up to 90 minutes to complete, with an additional 50 minutes for the child assessments. Furthermore, additional time was also required for locating and initial contact, plus time to complete the interviewer observations and interviewer remarks, for a total of 20 hours per completed case when all components were included. Three-quarters of the way into the field period, only 43 percent (N = 642) of the 1,500 cases had been completed. Among completed cases, multiple visits were required to complete the home visit. In 30 percent of the completed cases, more than one visit was required to collect the data. Usually, within this 30 percent, field staff needed two visits to complete the home visit, but occasionally required up to four home visits.

Interviewers also complained about the complexity of the ECLS-B protocol, including the administration of the BSID-II. The BSID-II is a complex assessment that requires knowing the steps for item administration, how to manipulate materials, and how to score the child's performance. The administration manual is geared to clinicians and professionally trained child development experts who must memorize item administration and scoring criteria in order to administer the BSID-II. Clearly, this was not a reasonable demand for field interviewers with no background in assessment or even child development.

As described in section 2.1.3, the ECLS-B had acquired the BSID-II materials used by MPR on the national Early Head Start Research and Evaluation Project. These materials included a flipbook of core items for the mental scale that were administered to all children and scoresheets for the core item scores. To make it easier for interviewers to keep track of the basal and ceiling rules, the basal items and the ceiling items were bundled into a supplementary flipbook. Based on the total score tallied on the scoresheet, the interviewer determined whether the basal items or the ceiling items needed to be administered.

Training videos used by MPR were also obtained and used as models. The age of the ECLS-B children at BSID-II testing differed, however, and Westat produced its own BSID-II training video for the 9-month data collection. In addition, Westat followed MPR's approach to certification of interviewers' proficiency in BSID-II administration. However, many of MPR's interviewers on the

national EHS evaluation study were professionally trained (e.g., graduate students in child development), and some were paraprofessionals (such as former kindergarten teachers). Although the ECLS-B hired only experienced interviewers, none had any experience or background in child assessment.

In addition, the ECLS-B used both the mental scale and the motor scale items, making it a more difficult task than in the Early Head Start evaluation study. Interviewers found the basal and ceiling rules confusing and were also unsettled by the administration flexibility required by the BSID-II. This confusion was evidenced by the high rates of errors found on the BSID-II scoresheets during quality control review during the fall 1999 field test. A high percentage of interviewers failed to apply the basal item set or the ceiling item set when it was necessary. There were also high percentages of missing items that indicated that interviewers did not know where to begin, or to stop, administering items. In addition, questions being sent in from the field indicated that many interviewers did not understand the intent of many of the items and therefore the accuracy of scoring was uncertain.

Finally, perhaps the most compelling evidence for the burdensome quality of the BSID-II was that administration time averaged 43 minutes for the BSID-II itself. A total of 30 minutes out of the 90-minute home visit had been allotted for the entire three-part direct child assessment component.

2.1.5 Development of the Bayley Short Form–Research Edition

To address problems noted in the field test related to the administration of the BSID-II, steps were taken to reduce burden. NCES and Westat conferred with members of the ECLS-B Technical Review Panel at the semiannual meeting about the September 1999 field test production problems. The Technical Review Panel was made up of representatives from the research, policymaking, and university communities who contributed to the design of the ECLS-B by reviewing research priorities and technical issues, questionnaires, and operational practices. NCES and Westat presented the following alternatives to the BSID-II: replacing it with the Bayley Neurodevelopmental Screener; using a parent report measure such as the (Minnesota) Child Development Inventory; dropping the BSID-II at 9 months altogether; or just administering the BSID-II motor scale only, which at 9 months tends to have greater predictive validity than the mental scale. The consensus that arose from that Technical Review Panel was that a direct assessment of children's developmental status at 9 months 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. Specifically, the Technical Review Panel members 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. This approach was used successfully to develop the ECLS-K assessment battery. (Please see *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05), available on-line at http://nces.ed.gov/ecls.)

2.1.5.1 Expert Panel Advice

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. Selection of items on the basis of face validity or the simplicity of materials required would not be adequate to ensure a psychometrically sound instrument. Therefore, the first step towards developing an abbreviated BSID-II was to form a panel of consultants composed of experts in educational and developmental assessment, psychometrics, and IRT analysis to make recommendations for the abbreviated Bayley. This panel of experts formed the assessment workgroup and included Dr. Don Rock, from the Educational Testing Service, who has extensive experience developing adaptive tests and also served in this expert capacity on the ECLS-K. A second work group member, Dr. Kathleen Matula, is an expert in early child assessment and had been project manager for the restandardization of the BSID-II while at The Psychological Corporation. Dr. Barbara Wasik of Johns Hopkins University is a researcher in developmental and educational psychology with extensive experience assessing cognitive development in low socioeconomic status and language minority samples. Dr. Kathleen Williams, then of American Guidance Systems, is a psychometrician with extensive experience developing standardized assessments. These assessment workgroup members concurred with the intention to use IRT techniques to identify the strongest items to be included and reviewed Westat's plan for the development of the short form of the BSID-II and provided comments about the results obtained from the field test and national data collections. They also recommended that the goal should be to obtain standard errors of 0.4 or less for scale scores on both the mental and motor scales, which should lead to a forecast reliability of 0.80 or greater on a 0 – 1 scale. The consultants also agreed that a two-parameter logistic (2-PL) IRT model[2] would be appropriate, the 2 parameters representing item difficulty level and item discrimination power. The advantages of IRT modeling are discussed in section 2.1.8. The 3-parameter logistic (3-PL) model

[2] For further information about IRT analysis in general and the 2-PL and 3-PL models, please see S. Embretson, S. Reise, and S. Paul. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum.

was not appropriate in this case, the consultants concurred, because the third parameter is a correction for guessing, and infants are not presumed to guess at the correct response to an item.

Additionally, members were consulted to ensure the quality of the data collection and implementation of the measures. In addition to her participation in the assessment workgroup, Dr. Kathleen Matula also served as an expert consultant for the BSF-R. She was consulted extensively about the administration and scoring of the BSID-II items that were included in the BSF-R. Her involvement focused on four areas. First, she provided clarifications about any ambiguities in the administration steps and the scoring of the items, for example the number of trials allows for "Puts 1 cube in cup" and "puts 3 cubes in cup," and whether the two items could be combined into a single administration. In addition, the BSF-R section of the Child Activity Booklet was sent to Dr. Matula for her review to make sure that all items were accurately represented. She also reviewed the accuracy of the BSF-R training videotape that was produced by Westat to ensure consistency in the training of 240 interviewers for the national study. Finally, Westat child development staff who would be trainers at the national training, were each videotaped administering the BSF-R to an infant. Their videotapes were sent to Dr. Matula, who reviewed the videotapes to make sure that these individuals administered and scored the BSF-R appropriately and maintained good rapport with the child being assessed.

2.1.5.2 Permission to Develop an Abbreviated BSID-II

The second step was to contact the Psychological Corporation, publisher of the BSID-II, explain the project's goals, and request access to the standardization data set in order to conduct the IRT analyses. The Psychological Corporation has permitted the development of an abbreviated BSID-II for research purposes and in its licensing agreement with the Department of Education has given its permission to call it the "Bayley Short Form–Research Edition" (BSF-R).

2.1.5.3 Creating the Bayley Short Form–Research Edition: Psychometric Rigor and Administrative Ease

Work towards developing a shortened Bayley was guided by two considerations: psychometric rigor and administrative ease. Psychometric rigor was obtained through IRT analyses to ensure that the psychometrically strongest items were included. These analyses are described below.

To oversimplify a bit, one function of IRT analyses is to line up all the items in question according to their ability level. Ideally, the items will line up at evenly spaced intervals all across the ability range. Using the publisher's standardization data set, the ability distribution of items appropriate for the 9-month range was identified and then extended at each end to take into account any infants born prematurely and those infants 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 9-month data collection BSF-R mental scale ranges from –3.43 to 1.01 population standard deviations (where the 12-month population[3] has a mean of 0 and a standard deviation of 1), which corresponds to an item age range from 4 to 19 months. The ability distribution for the 9-month BSF-R motor scale ranges from –3.41 to 0.98 standard deviations, with an item age range from 3 to 22 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, although sometimes even lower values were necessary if there were no better items in the appropriate range of ability. This resulted in a large pool of candidate items for the BSF-R, still too many to be practical for household administration. However, several more items were deleted on the basis of redundancy of coverage—if two items represented the same construct, say "object permanence," 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 were reasonable to administer in a field setting by field interviewers. Included items must also have had relatively objective scoring criteria. Administrative selection criteria were formulated to complement the IRT analytic criteria, as described in the following.

Administration difficulty. Items that were difficult to administer (e.g., because of procedural complexity or unwieldy materials) were targeted for deletion. For example, "Puts peg in pegboard" requires at least 2 minutes and 10 seconds to administer. In addition, the pegs typically roll onto the floor and need to be retrieved, which takes even more time.

Minimal materials. Minimizing the number of materials needed was another important consideration. The ECLS-B interviewers have about 25 pounds of equipment to carry, including laptops,

[3] 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.

physical measurement equipment, and video cameras. Anything that could be done to reduce the number of BSF-R materials was desirable.

Objectivity of scoring. It was also desirable to exclude items with difficult or subjective scoring criteria, e.g., "Listens selectively to two familiar words," or "Purposely removes lid from box to find toy." In general, the ECLS-B interviewers had no training in child development, early childhood education, or testing. It was difficult for them to make inferences about infant intentionality during responding. Therefore, item scoring criteria needed to be as objective as possible so that interviewers would know what to look for.

Maximize "twofers." In the BSID-II, it is sometimes the case that multiple scores can be obtained from one administration. For example, "Puts one block in cup," "Puts three blocks in cup," or "Puts nine blocks in cup" can all be scored from one administration. To the extent that it was possible, given the constraints of 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 this introduces the problem of interdependence of items. Analytically, after the completion of the data collection, one of the interdependent items was deleted from the equating, which is discussed in more detail below.

Breadth of content. An additional goal was to maintain as much of the content of the BSID-II mental items as possible. The BSID as originally constructed is rather atheoretical; it is based on the author's observations of children's abilities and incorporates successful items from other assessments, such as the Gesell Developmental Schedules (Gesell 1949). However, a factor study of the BSID-II has shown that up to about 24 months of age there is only a single main factor, which can be described as general "ability," with some indications of subfactors, such as object permanence. Although breadth of content was a goal at 9 months, it becomes more important at the later ages.

2.1.6 Age Set Structure of the BSID-II and Developing an Adaptive Testing Strategy

The BSID-II includes 178 mental and 111 motor items designed for infants between 1 and 42 months of age. To avoid frustrating a child with items that are inappropriate for his age, specific age item sets have been recommended by the publisher, together with basal and ceiling rules to determine whether it is necessary to test outside the range of the designated age item set. 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 are required, all of the items in the supplementary item set are to be administered. Since adjacent item sets contain overlapping items, this usually requires administering four to ten items for each additional 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. The objective is to develop a core item set that is appropriate for most of the infants in the target age group. The raw score total for these core items can then be used to determine whether any specific infant should be administered additional basal or ceiling item sets. Indeed, this adaptive strategy closely parallels the standard procedures of administration recommended by the publisher of the BSID-II. In those cases in which an additional set is required (basal or ceiling), all the items in that additional set are administered.

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 infants in a clinical setting. For 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 highly discriminating items in each of the reduced item sets helps ensure measurement precision across the full range of the target population ability distribution.

2.1.7 Core-Basal-Ceiling Set Structure of the BSF-R

On the basis of the IRT analyses, review of the item characteristic curves (ICCs), and implementation of the above administrative selection criteria, the content of the 9-month BSF-R was determined. IRT analyses were also used to determine the structure of the core set of items that were to be administered to all children, the basal set of items that were to be administered to children who did not receive credit for many items, and the ceiling set of items that were to be administered to children who received credit for most of the items. The basal and ceiling sets were intended to provide additional information about children's performance when the core set alone did not provide a "good enough" estimate. Exhibits 1 and 2 summarize the items included in the mental and motor subscales of the BSF-R and the contents of the core, basal and ceiling sets.

The core set of the mental scale has 13 items that require only 9 administrations. These items are administered to all the children in the ECLS-B. On the basis of their performance on the core items, the interviewer may need to administer a supplementary set of items. Specifically, if the child receives credit (C) for 3 or fewer of the 13 core items, then the interviewer administers the set of basal items. If the child receives credit for 9 or more of the core items, then the interviewer administers the set of ceiling items. The criteria determining whether the basal set or the ceiling set needed to be administered were set using IRT analyses. Based on the normal distribution, it was expected that children greater than one standard deviation below the age-group mean would be administered the basal items and children greater than one standard deviation above the mean would be administered the ceiling items. This implies that, based on the standardization data set, close to 68 percent of children would need the core items only, with 16 percent needing the basal sets and another 16 percent the ceiling set.

In exhibit 1 it can be seen that for the mental basal set, only one item needed to be specially administered and all the rest of the mental basal item scores were obtained during the administration of the core items—almost all the mental basal items are "twofers." For the ceiling items, there were 6 administrations and 9 item scores, the extra 3 being obtained during the core item administrations.

For the motor scale, summarized in exhibit 2, the BSF-R core item set has 5 administrations that yield 14 item scores, with an additional 4 administrations in the basal set and 7 other basal scores being obtained either by observation or during the administration of the core items. And there are 5 items in the ceiling set that are administered with another 5 being obtained by observation during the core set for a total of 10 scores. The last column presents the age range of each item. The BSF-R items range in age from 4 to 19 months on the mental scale and from 3 to 22 months on the motor scale.

Exhibit 1. 9-month BSF-R mental scale items and set structure, by various characteristics: 2001–02

Material	Item number	Number of administrations[1]	Item description	Item age range[2] [months]
Basal item set (9 scores)		1		
Cube	45	Observe	Picks up cube	4-5
Red ring	48	Observe	Plays with string	4-5
Pellet	51	Observe	Regards pellet	4-6
Any toy	52	Observe	Bangs in play	5-6
Cubes	53	Observe	Reaches for second cube	5-6
Cup and rabbit	55	1	Lifts inverted cup	5-7
Cubes	57	Observe	Picks up cube deftly	5-7
Cubes	58	Observe	Retains 2 cubes for 3 seconds	5-7
None needed	61	Observe	Vocalizes 3 different vowel sounds	5-8
Core item set (13 scores)		9		
Bell	59	1	Manipulates bell, interested in details	5-8
	66		Rings bell purposely	5-11
Red ring	62	1	Pulls string to secure ring	5-8
Cubes	65	1	Retains 2 of 3 cubes for 3 seconds	5-10
Picture book	69	1	Looks at pictures in book	6-11
	73		Turns pages of book	6-12
None needed	71	Observe	Repeats vowel-consonant combinations	6-12
Blue box and beads	72	1	Looks for contents of box	7-12
None needed	76	Observe	Jabbers expressively/says at least one word	8-12
Little red car	77	1	Pushes car	8-12
Pegboard	79	1	Fingers holes in pegboard	8-13
None needed	81	1	Responds to spoken request	8-13
Cubes and cup	86	1	Puts 3 cubes in cup	9-13
Ceiling item set (9 scores)		6		
Blue box and beads	89	1	Puts 6 beads in box	11-16
Crayon and paper	91	1	Scribbles spontaneously	11-16
Cubes and cup	95	Observe	Puts 9 cubes in cup	12-16
Stimulus card	99	1	Points to 2 pictures	12-19
None needed	100	Observe	Uses 2 words appropriately	13-19
None needed	101	1	Shows shoe, clothing or object	13-19
Cups and rabbit	102	1	Retrieves toy—visible displacement	13-19
Rod and toy	104	1	Uses rod to attain toy	13-19
None needed	106	Observe	Uses words to make wants known	14-19

[1] An administration is defined as the structured presentation of stimulus material to obtain the child's response. Sometimes one administration can yield multiple scores, e.g., presentation of the sugar pellet yields scores for motor item number 32, 41, and 49. Sometimes behaviors only need to be observed for a score to be given, e.g., motor item 65, Squats briefly, or, if not spontaneously performed can be simply elicited, e.g., by putting a ball on the floor to observe squatting. Multiple items scored from one administration only count as one administration. 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 age item sets in which the item is included. Thus, an age range of 5-10 months means the item is included in the 5-month through the 10-month age item sets of the original BSID-II.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) Nine Month Data Collection, 2001–02.

Exhibit 2. 9-month BSF-R motor scale items and set structure, by various characteristics: 2001–02

Material	Item number	Number of administrations[1]	Item description	Item age range[2] [months]
Basal item set (12 scores)		4		
None needed		1		
Orange rod	25	1	Shifts weight on arms	3-5
None needed	29		Uses whole hand to grasp rod	3-6
	22		Sits alone with slight support for	3-4
	28	1	10 seconds	
	34		Sits alone momentarily (2 sec)	3-6
	36		Sits alone for 30 seconds	4-6
Bell	26	1	Sits alone steadily	5-7
	38		Turns from back to side	3-5
Any toy	30	Observe	Turns to back from stomach	5-7
Cube	31	Observe	Reaches unilaterally	4-6
	32	Observe	Uses partial thumb opposition to grasp cube	3-6
Sugar pellet			Attempts to secure pellet	4-6
Core item set (13 scores)		7		
Sugar pellet	41	1	Uses whole hand to grasp pellet	6-7
None needed	45	1	Pulls to standing	6-8
Sugar pellet	49	1	Uses partial thumb opp., grasp pellet	7-9
Squeaky toy	51	1	Moves from sitting to creeping	7-10
	43		Moves forward, prewalking movement	9-13
Pencil	58	1	Grasps pencil at farthest end	8-12
Squeaky toy	52	1	Raises self to standing	8-10
Squeaky toy	54		Walks sideways holding onto furniture	8-11
None needed	40		Early stepping movements	5-7
None needed	44		Supports weight momentarily	6-8
None needed	46	1 combined	Shifts weight while standing	6-8
None needed	53		Attempts to walk	8-10
None needed	60		Walks with help	8-12
None needed	61		Stands alone	9-13
Ceiling item set (10 scores)		5		
	59	1	Stands up I	8-12
None needed	68		Stands up II	11-19
None needed	62	1 (or by	Walks alone	9-13
None needed	63	observing)	Walks alone with good coordination	10-16
None needed	65		Squats briefly	11-16
Ball, if needed	67	Observe	Walks backward	11-19
Pull toy	70	1	Grasps pencil at middle	12-22
Pencil	71	Observe	Walks sideways	12-22
Pull toy	72	1	Stands on right foot with help	12-22
Squeaky toy	73	1	Stands on left foot with help	13-22
Squeaky toy				

[1] Each administration is defined as the structured presentation of the stimulus material to obtain the child's response. Multiple items scored with the same administration only count as one administration (including item trials). 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 for which the item is included. Thus, an age range of 5–10 months means the item is included in the 5-month through the 10-month item age sets of the original BSID-II.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine Month Data Collection, 2001–02.

2.1.8 Item Response Theory Item Calibrations of BSID-II Standardization Data Set

IRT principles were used to develop a BSF-R that is as comparable as possible with 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, estimates of item and ability parameters are sample free—that is, 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, that is, 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 reliable estimates. IRT offers the prospect of providing comparable scores that share the same scale metric found in publisher data.

The BSF-R was designed with IRT techniques to produce results that are as consistent as possible with those obtained using the BSID-II. There are four steps in this process, which can be summarized as follows:

- IRT calibration of the full complement of 178 mental and 111 motor items comprising the BSID-II using a 2-PL IRT model and the publisher BSID-II standardization data set.
- Consulting publisher IRT item difficulty and discrimination parameters to select optimal subsets of core, basal, and ceiling items for the BSF-R.
- Field testing of 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.
- Final BSF-R item calibrations, using the ECLS-B 9-month national data set, final IRT true score equating using the publisher test as the target, generation of ability estimates, and indices of infant development reported in publisher scale metrics.

The assessment of individual differences in general (e.g., personality, intelligence) employs measures subjected to rigorous evaluation based on decades of psychometrics research. The testing movement has made permanent contributions to statistical methodology by producing a specialized branch of applied statistics known as psychometrics, which has come to dominate thinking about the reliability and validity of psychological measurement. Psychometric techniques are widely used with measures of cognitive skill and school achievement, but have also been applied to measures of personality and social behavior. The assessment of measures of infant development uses an analytic strategy that is prominent in modern psychometrics: item response theory (IRT).

IRT has been developed to represent what happens 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.

A function known as an item characteristic curve (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 progressively 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 determined 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.

Indeed, one of the congenial features 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 backwards and infer the examinee's level of ability.

The item characteristic curve (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, how the item is scored. The BSID-II is based on a series of items representing infant behavior. Instead of answering items on a test, as older children do at school, infant behavior is observed on a series of specific tasks presented by an examiner. Item responses are based on the examiner's perception of infant behavior as he or she attempts to undertake each task.

The examiner records whether or not the infant 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 infant is presented a task that he is asked 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.

Examinee observations of infant 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 infant. 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 dislocates the ICC left and right 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 at each ability level. For example, the item MEN066, Rings bell purposely, has an ability parameter of -2.39 and a discrimination parameter of 1.55, whereas MEN095 has an ability parameter of 0.70 and a discrimination parameter of 0.90. Parameter estimation is referred to as "item calibration" and involves fitting the ICCs to the actual item responses. Parameter estimates are selected that maximize the likelihood of item responses across all ability levels for the sample as a whole. The likelihood of ability estimates (θ) are calculated concurrently as part of the item parameter estimation cycle. Several iterations of maximum likelihood estimation 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 are used in an analysis where substantive issues of infant development can then be examined.

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 infants under clinical conditions. The combined sample includes a standardization sample of 1,700 observations of normal infants, arranged in 17 age groups, ranging from 1 month to 42 months of age, by month from 1 to 6 months, bimonthly from 6 to 12 months, trimonthly from 12 to 30 months, and semiannually from 30 to 42 months, and 1,239 additional observations. (Hence, there is no standardization data for the 9-month age group per se. This data is interpolated based on the 8- and 10-month age sets.). 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 infants of different ages. The standardization sample contained 100 observations for each of the 17 selected age groups (table 3).

Table 3. Raw scores and index scores means and standard deviations on mental and motor scales for the BSID-II standardization sample, by age group: 1993

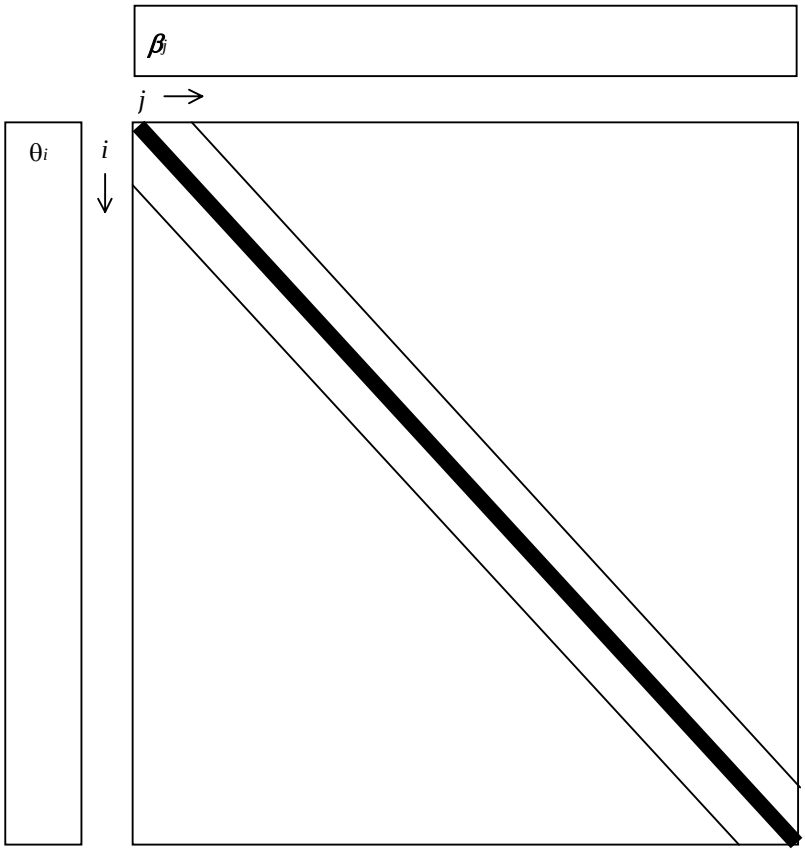
Months of age	Sample *N*	Mental scale				Motor scale			
		Raw score		Index score		Raw score		Index score	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	100	15.3	9.4	101.8	18.2	11.7	3.9	101.5	13.5
2	100	27.4	7.2	99.8	14.7	16.5	5.4	100.0	15.0
3	100	33.5	7.9	100.0	15.7	25.0	7.0	99.4	19.6
4	100	44.5	7.7	99.9	15.4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

SOURCE: Standardization data set 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 month 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 overlap with and provide 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 1 represents the core item sets recommended for adjacent age groups, with limited overlap in basal and ceiling items linking adjacent core item sets, where i = rows of individuals sorted by individual ability (θi), and j = columns of items sorted by item difficulty (βj).

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, 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 data set, 8.9 percent of all children were administered basal items, while 14.1 percent received ceiling items.

Figure 1. Schematic representation of publisher data, IRT ability estimates θ_i, and item parameters β_j



NOTE: *i* = rows of individuals sorted by individual ability (θi) and *j* = columns of items sorted by item difficulty (βj).
SOURCE: Standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

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. 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 month 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

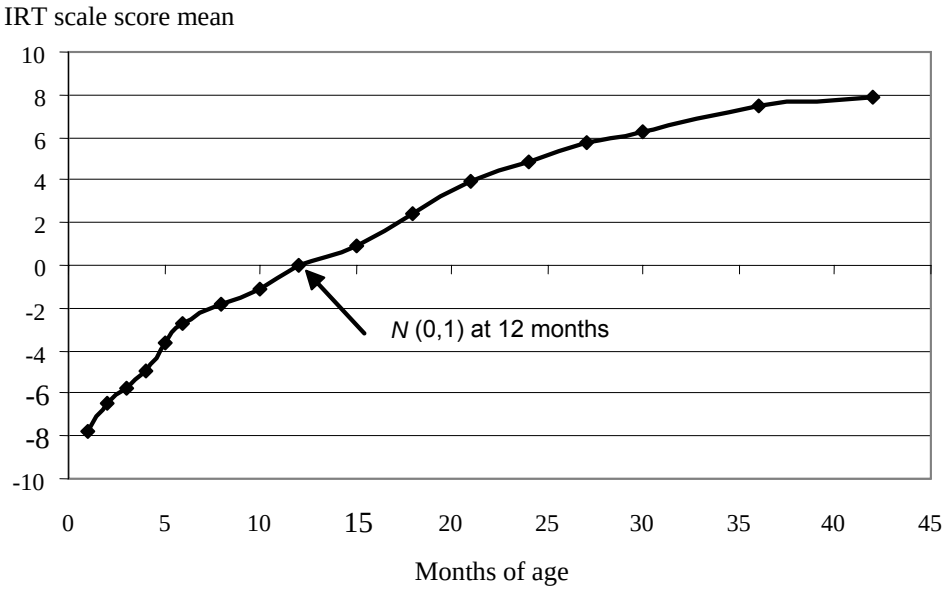
assuming a standard normal $N(0,1)$ distribution for latent ability, with population mean $\mu = 0$ and standard deviation $\sigma = 1$.

The 12-month 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 be the center and 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 multi-group 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 outwards 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 infancy is quite explosive, the resulting development scales span many population standard deviations between 1 month and 42 months of age. For the mental scale, estimated population means for the different age groups range between $-8\sigma < \theta < 8\sigma$, with population standard deviation $\sigma = 1$, as shown in figure 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 suggests that early infant 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.

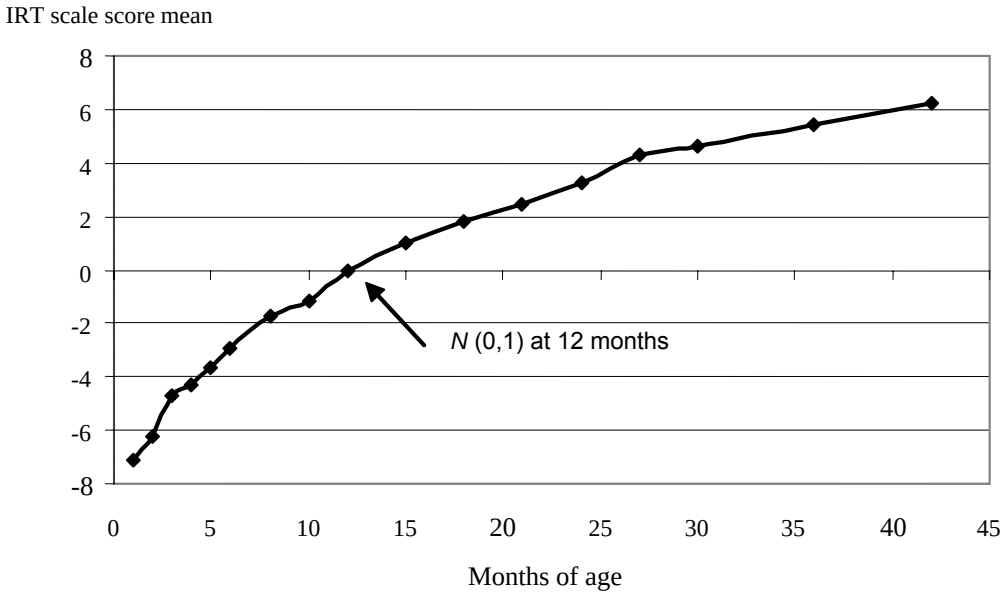
Figure 2. Polynomial fit showing BSID-II mental scale score means by age: IRT 2-PL item calibrations using BSID-II standardization sample: 1993



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

For the motor scale, population means range between $-7\sigma < \theta < 6\sigma$, with population standard deviation $\sigma = 1$, as shown in figure 3. By working outwards 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 month and 42 months of age. The age-specific latent ability distributions have standard deviations that are nearly equal to one, 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 3. Polynomial fit showing BSID-II motor scale score means by age: IRT 2-PL item calibrations using BSID-II standardization sample: 1993

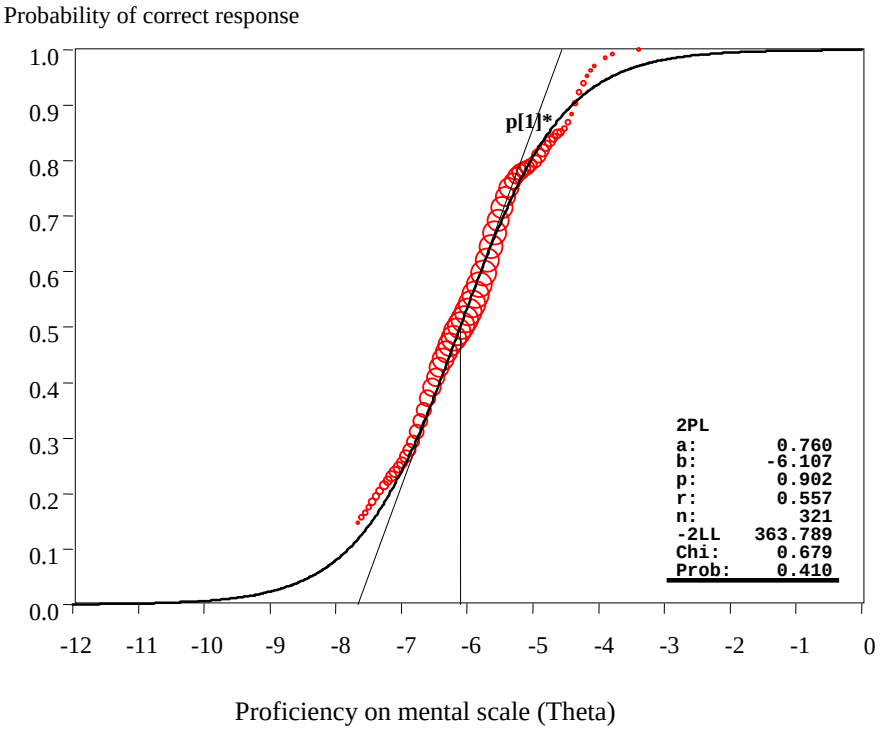


SOURCE: Standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Concurrent estimation yields item calibrations similar to that shown as an example in figure 4. The numbering of BSID-II items is intended to reflect the item's relative difficulty. Mental item number 28 (MEN028) is the 28th item among 178 mental scale items, implying that it is one of the easiest items in the BSID-II and is recommended for infants 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, implying that it is more difficult than 27 items and less difficult than 150 items. A raw score of 28 points falls between the standardization sample means for infants 2 and 3 months of age.

The ICC for this item is rising opposite scale values on the x-axis in the vicinity of $\theta = -6.1$. Accordingly, the IRT ability parameter for this item is $b = -6.107$, reported in the box to the right of figure 4, and is represented by a vertical line rising to the midpoint inflection 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 shows that this is indeed the appropriate scale range for infants between 2 and 3 months of age.

Figure 4. Item characteristic curve (ICC) for item MEN028 (Displays Visual Preference) representing the probability of a correct response: IRT 2-PL item calibrations using BSID-II standardization sample: 1993



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

The IRT difficulty parameter reflects the 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 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 plenty 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 $r = .99$ for both the mental and motor item sets.

The box to the right in figure 4 reports 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 are relatively less steep.

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 data set than the Rasch model, which has only an item difficulty parameter and has no provision for items that vary in discrimination. Some BSID-II items are more discriminating than others.

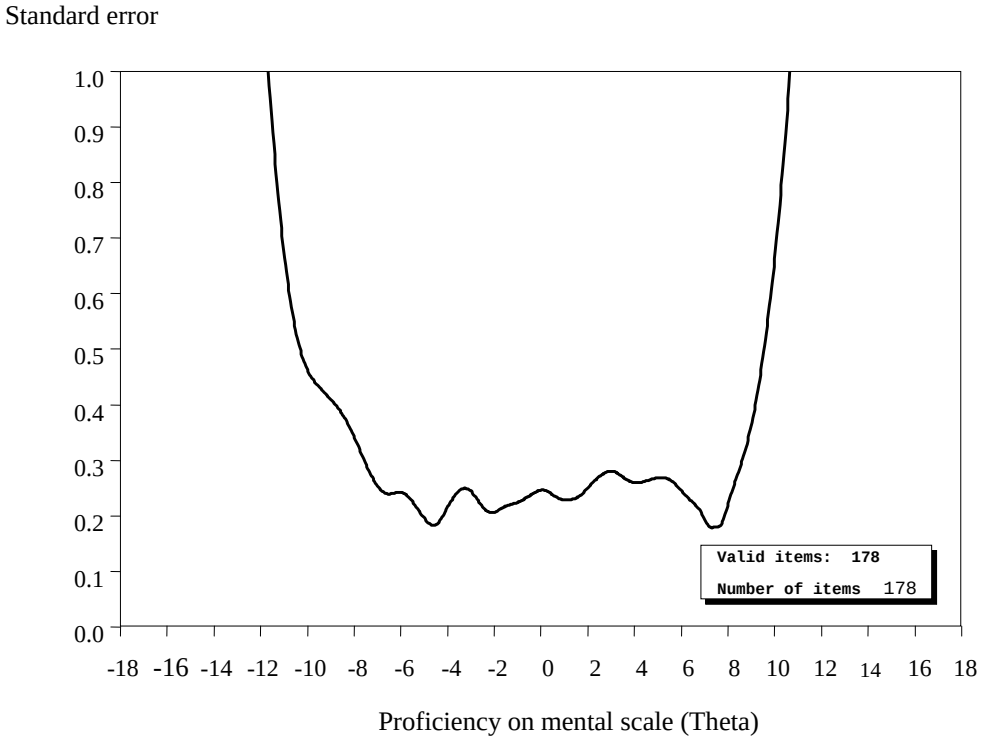
The circles in figure 4 are drawn to scale to represent the number of observations in the calibration data set and reflect response probabilities assuming that the 2-PL response model is appropriate. When the model fits the data, the circles will align with the ICC function. Visual inspection and chi-squared (χ^2) statistics suggest that there are perhaps a dozen or so mental items (6 percent or 11 items in 178) that are marginally represented by the 2-PL model. Although the quality of fit also varies for motor items, for the motor scale, it appears that virtually all of the items fit the model. With only minor shortcomings in terms of fit, all 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. Information is higher when the value of the *a* parameter is more expressive and when item difficulty *b* coincides with examinee ability (θ). In other words, items with considerable power of discrimination, at the appropriate level of difficulty for the examinee, convey the most information about his 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 infant subjects. 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. A greater number of items, 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 5.

Indeed, the standard error across most of the ability distribution is SE(θ) < 0.3, implying that the errors are less than a third of a population standard deviation across virtually all of the distribution that is relevant for infants between 1 month and 42 months of age.

Figure 5. Standard error of measurement for the BSID-II mental scale: IRT 2-PL item calibrations using BSID-II standardization sample: 1993

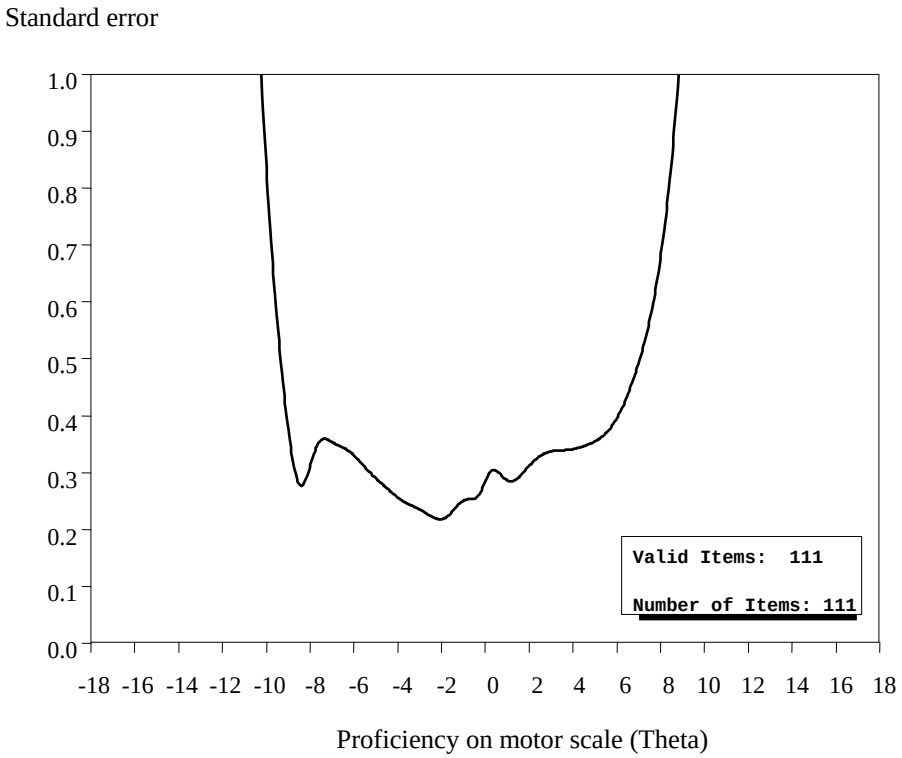


SOURCE: Standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

The standard error of measurement for the IRT motor scale is shown in figure 6. Precision 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 month 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} = .94 for the IRT mental scale and r_{xx} = .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-Richardson)

coefficients of internal consistency averaging r_{xx} = .88 for the mental scale and r_{xx} = 0.84 for the motor scale across all age groups.[4]

Figure 6. Standard error of measurement for the BSID-II motor scale: IRT 2-PL item calibrations using BSID-II standardization sample: 1993



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

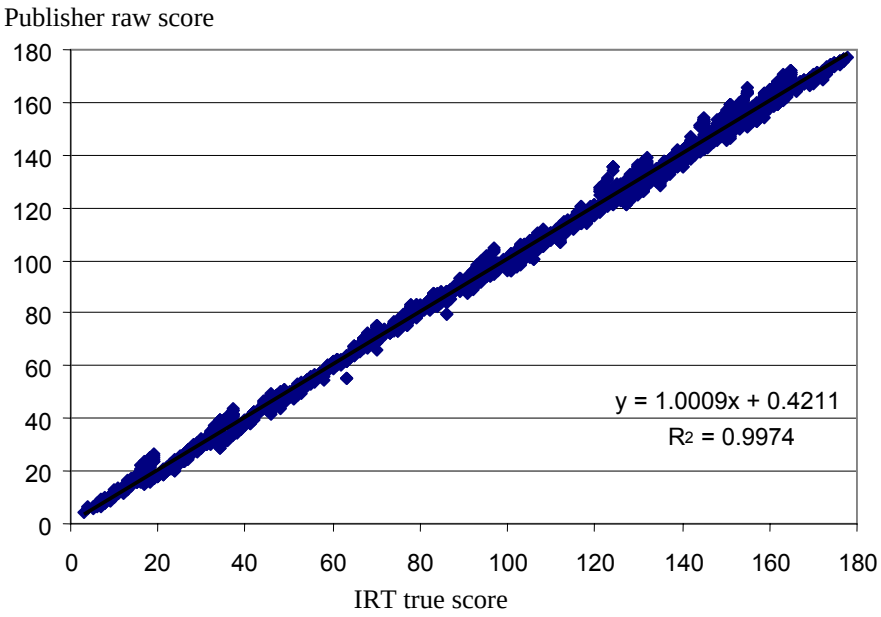
The objective of testing is to assign a score to an individual examinee that reflects the level of attainment of a skill. One approach to scoring is to give a point for each correct response and present the test outcome as an item-correct raw score. Indeed, this is the origin of the number-right raw 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 usually different from obtaining 14 right out of 50. A means must be found to enable item substitution and deletions without altering the scale metric used to

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

express test results. IRT has been developed to enable this flexibility. However, first it must be shown that IRT ability estimates θ can be reported using publisher raw score metric.

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

Figure 7. Relationship between IRT true score and publisher raw score for the mental scale: IRT 2-PL item calibrations using BSID-II standardization sample: 1993

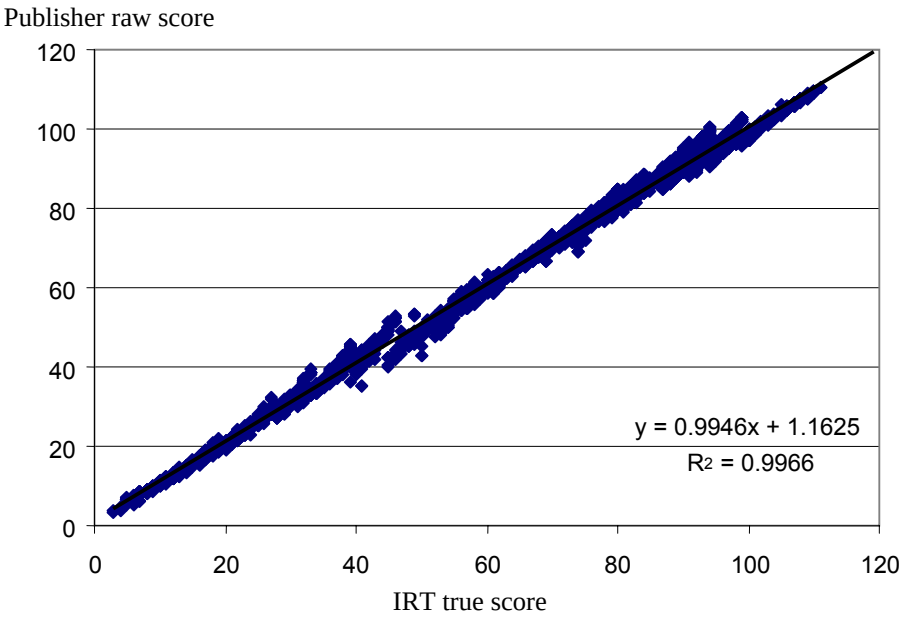


SOURCE: Standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

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 = .997). Figure 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 = .996). 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.[5]

Figure 8. Relationship between IRT true score and publisher raw score for the motor scale: IRT 2-PL item calibrations using BSID-II standardization sample: 1993



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

2.1.9 Selecting Item Sets for the Bayley Short Form–Research Edition

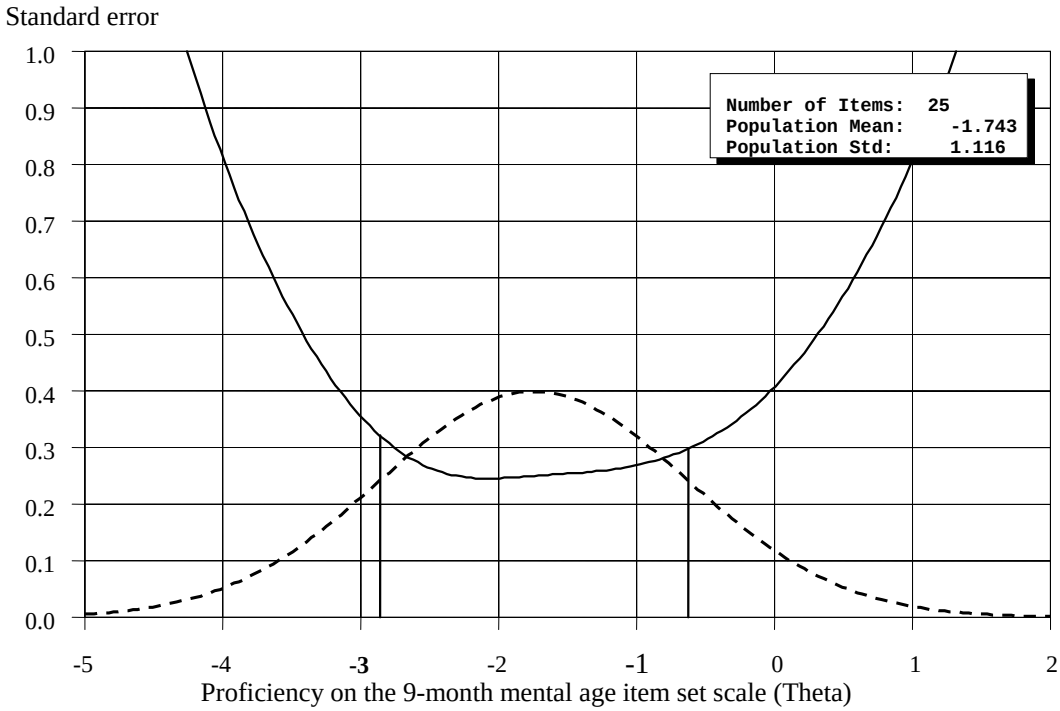
Once the 178 mental and 111 motor items have been calibrated using publisher data, it is possible to predict how subjects will respond to items before any test is taken into the field. Item parameters define an item response function representing the probability of a correct response by any examinee. 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 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.

[5] Standardized scores are reported by the publisher as "development index scores." In the ECLS-B, standardized scores are called "T-Scores."

In order to select reduced item sets for the BSF-R, the technical properties (i.e., difficulty and discrimination parameters) of items in the 9-month age item set recommended by the publisher were examined. The 9-month age group was selected as the target population in keeping with the base year data collection that was to take place when the children turned 9 months of age. There are no observations for this age group in the publisher standardization sample. However, interpolation of scale score values can be used to estimate the population mean and standard deviation for this age group, based on results obtained for the 17 age groups included in publisher data.

Figure 9 shows the respective target population ability distribution superimposed on a graph of the standard error of measurement SE(θ) obtained for the 25 items in the age item set recommended by the publisher for 9-month-old infants, superimposed on a graph of the respective target population ability distribution. For reference purposes, the 9-month frequency distribution $N(\mu, \sigma)$ appears in the background and is represented by a dashed line. The vertical lines connecting the graph with the abscissa are placed one population standard deviation to the left and right of the mean for the 9-month population, $\mu \pm \sigma$. Approximately 68 percent of the infant population falls within $\mu \pm \sigma$.

Figure 9. Standard error of measurement for the 9-month mental age item set: IRT 2-PL item calibrations using BSID-II standardization sample: 1993



SOURCE: Standardization data set 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 25 items in the age item set recommended by the publisher afford considerable measurement precision for 9-month-old infants within the limits of $\mu \pm \sigma$. Moving outwards 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 recommended 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 were to be administered until the basal rule or the ceiling rule was satisfied.[6]

For use in the ECLS-B, the BSF-R was designed to reduce administration time without compromising the quality of infant development data. 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. At an early stage in the development of

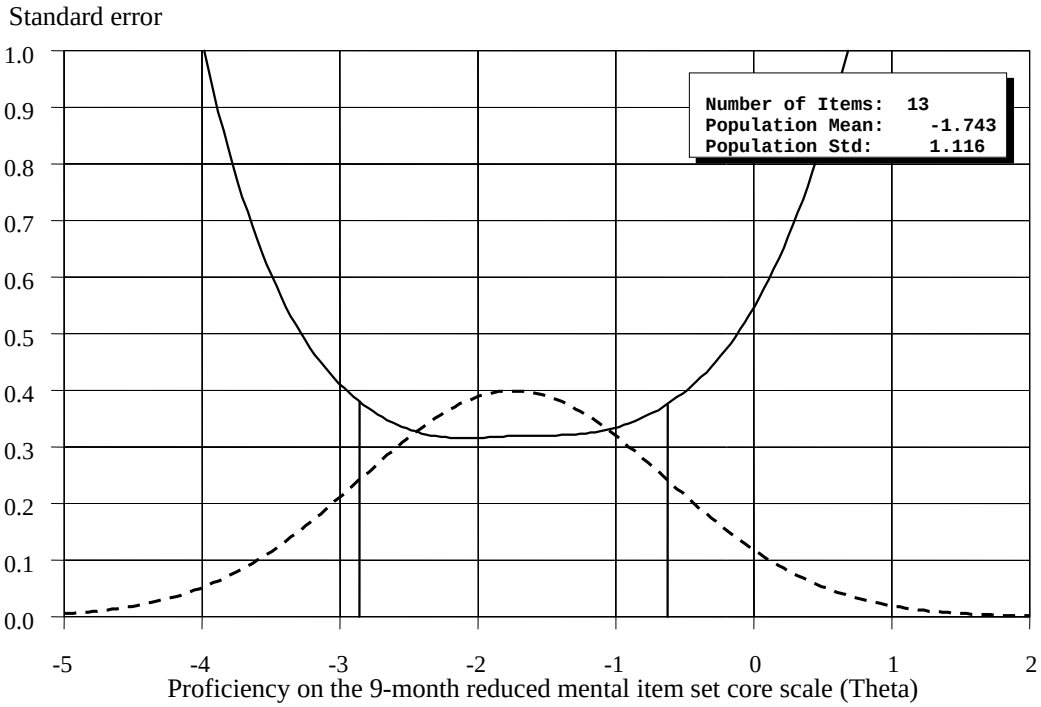
[6] This compares with the administration of only a single basal item set or a single ceiling item set on the BSF-R.

BSF-R, it was decided to construct a short form that would yield standard errors of measurement in the vicinity of SE(θ)= 0.4 across the target population ability distribution, extending well out into the tails. This corresponds with a reliability coefficient of approximately r_{xx} = .8.

The selection of reduced item sets for the BSF-R began by examining the most highly discriminating items available in the range of difficulty appropriate for 9-month-old infants. For the core item set, the range of difficulty is approximately $\mu \pm \sigma$. Within this general range of difficulty, priority of selection was given to the most discriminating items, those where item discrimination parameter values exceed $a > 0.7$. These criteria may be compromised, depending on the availability of appropriate items in the item pool. 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 infants in an age-specific target population. The approach used in the BSF-R is illustrated beginning with the standard errors for the 9-month mental core item set presented in figure 10. The BSID-II contains 24 items in the 9-month age set. Of these, a set of 13 items satisfied all of the above criteria and was used to construct a hypothetical mental scale. Items calibrated with publisher data could then be used to estimate the new scale's standard error of measurement across the full range of ability. A comparison of this figure with figure 9 shows that the new 13-item scale was not as precise as the 25-item scale based on the publisher's recommended age item set. The reduced item set had errors in the middle of the distribution just above SE(θ) = 0.31, whereas figure 9 shows errors closer to SE(θ) = 0.25. However, the reduced item set afforded standard errors that met or exceeded the objective SE(θ) = 0.4 over the range $\mu \pm \sigma$.

Figure 10. Standard error of measurement for the BSID-II 9-month mental reduced core item set: IRT 2-PL item calibrations using BSID-II standardization sample: 1993



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

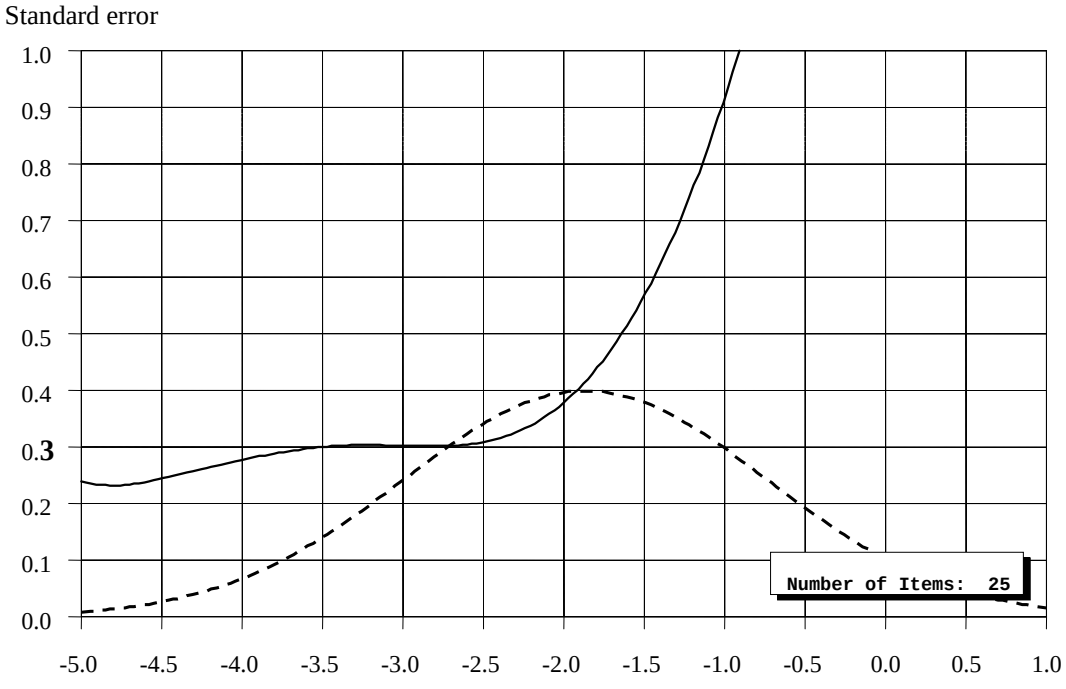
The new scale yielded satisfactory precision across the central part of the target population's latent distribution, where approximately 68 percent of the infants were to be found. This suggested that the assessment of many infants would require no more than 13 items. Outside the range $\mu \pm \sigma$, appropriate basal and ceiling items would have to be administered so that the objective $SE(\theta) = 0.4$ would be satisfied at the tails of the distribution. This suggested a strategy for adaptive testing that would yield appropriate measures for all of the infants in an age-group while still reducing the burden of fieldwork. An adaptive testing strategy offered advantages in terms of efficiency and provided enhanced precision for individuals with exceptional levels of ability.

The majority of infants (approximately 68 percent) needed to be administered only the reduced core set of 13 items. Results obtained with these items were 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 be administered to only those infants who required them. A routing test was thus used to administer two second-stage tests, both of

which were designed for extreme levels of ability. In order for this adaptive testing strategy to work properly, suitable reduced core, basal, and ceiling item sets needed to be found. Also, a decision rule governing the application of basal and ceiling items, based on results obtained with the initial core set, was needed.

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 5-month set (figure 11) to be a likely source of highly discriminating basal items, appropriate for the 9-month population below μ-$\tilde{\sigma}$. These items, all within the 5-month age set, ranged in difficulty from 4 to 8 months. 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 basal set of 9 items was selected. These items would only be administered as a complement to the BSF-R core item set. Consequently, it was not necessary to examine the technical properties of a hypothetical scale comprising basal items alone.

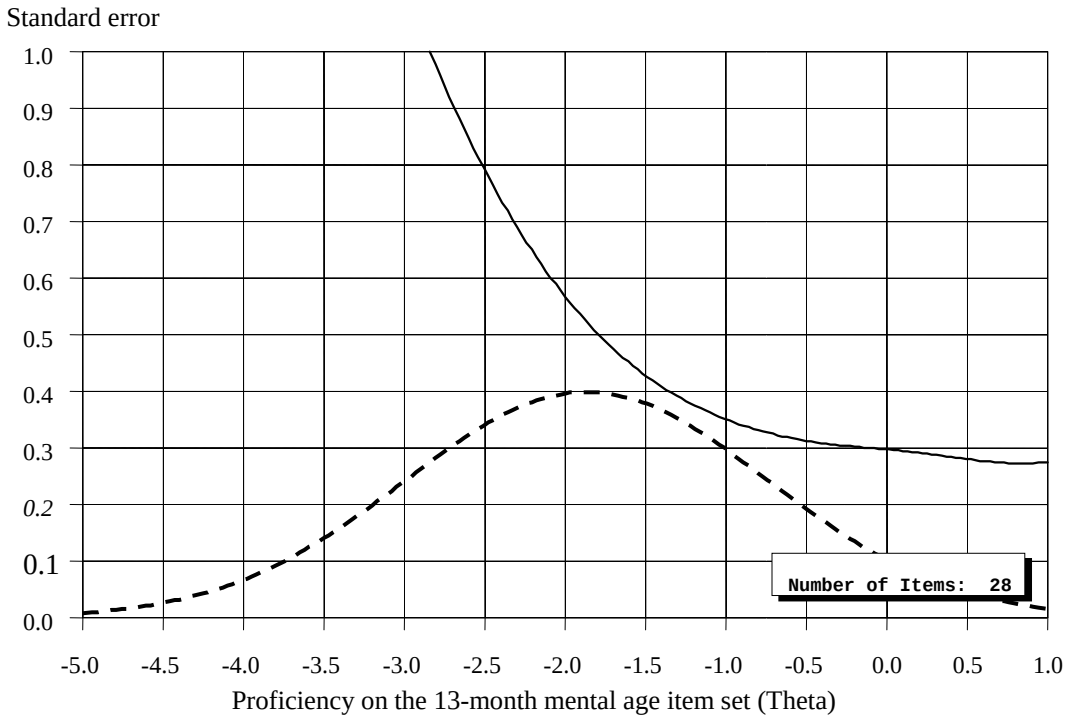
Figure 11. Standard error of measurement for the BSID-II 5-month mental age item set: IRT 2-PL item calibrations using publisher data



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

At the other end of the ability distribution, items needed to be found for a reduced ceiling item set (figure 12). The BSID-II item pool was once again examined, and the 13-month item set was found to be an appropriate source of ceiling items. Although these items were all within the 13-month age set, they ranged in difficulty from the 11- through 19-month age sets. Item selection criteria were again taken into account and 9 items were selected for the reduced ceiling item set. This ceiling item set would only be administered as necessary following the core items.

Figure 12. Standard error of measurement for the BSID-II 13-month mental age item set: IRT 2-PL item calibrations using publisher data

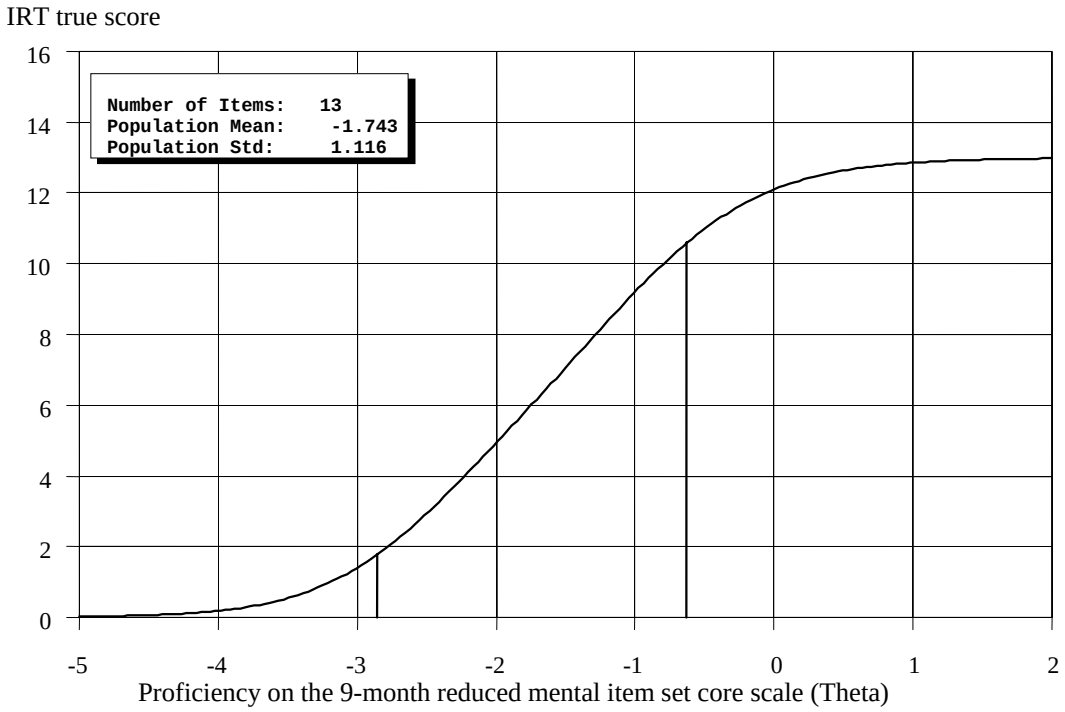


SOURCE: Standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

For the adaptive testing strategy to work properly, basal and ceiling decision rules needed to be devised. The decision rules are sufficiently simple so that they could easily be followed in the field and were based on counting the number of correct responses. The number-right raw score was based on this same counting procedure. 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 9-month mental reduced core item set is shown in figure 13.

Figure 13. IRT True Scores for the 9-month BSF-R mental core item set: IRT 2-PL item calibrations using BSID-II standardization sample: 1993



SOURCE: Standardization data set 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 2 at the lower end and 11 at the high end of this range. The decision rule that was actually

defined for the BSF-R mental scale at the low end was that 1, 2, or 3 points on the core item set would require administration of the reduced basal item set. At the high end, the rule was that 11, 12, or 13 points on the core item set would require administration of the reduced ceiling item set.

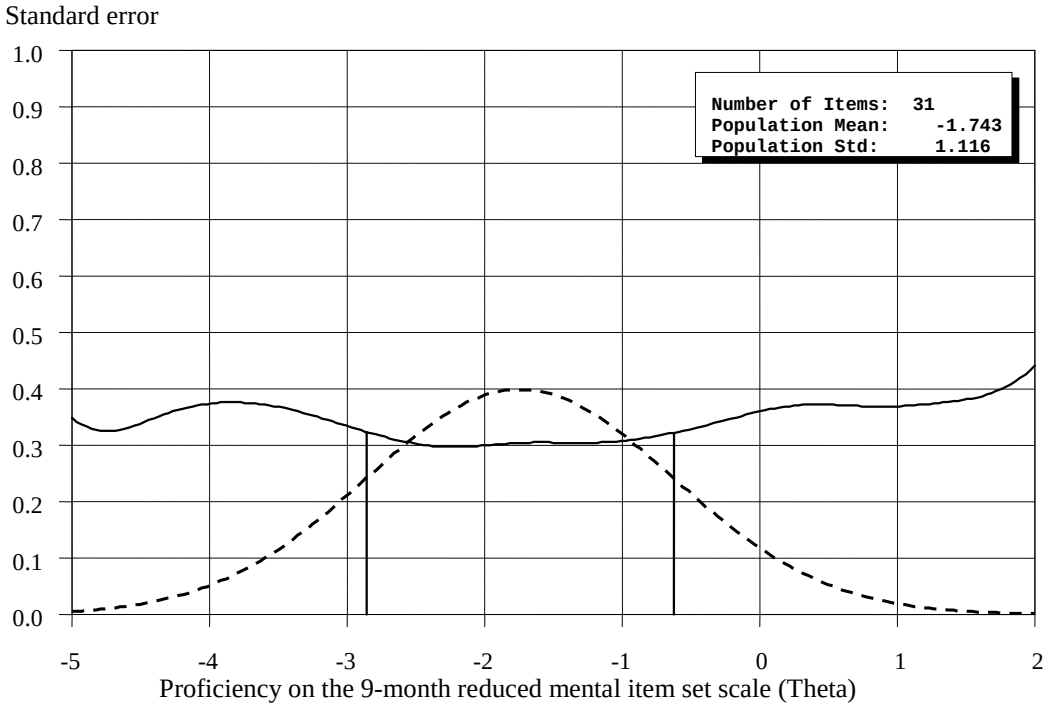
The 9 basal items, 13 core items, and 9 ceiling items contributed a total of 31 items to the BSF-R mental scale. A child would never be administered all of these items. Neither would the basal or ceiling items be administered by themselves, but rather only after first administering the core item set. Consequently, a child could be administered either 13 or 22 items, depending on whether the core items were sufficient or whether the basal or ceiling items were also required. Approximately 68 percent of the target population would receive only the 13 core items. Another 32 percent would 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 received 13 items whereas the remaining 32 percent received 22 items, so that on average across the entire sample, 20 items were administered to each child. Consequently, the expected average is (68 percent x 13) + (32 percent x 22) = 15.88 mental items administered on average.

Figure 14 shows standard errors for the 9-month BSF-R mental scale based on item calibrations obtained with publisher data. Although the figure is based on all 31 items, it is at least approximately correct for the core, basal, and ceiling item combinations that will be 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 were administered only the reduced core item set would have somewhat larger errors than those depicted in the figure if their abilities lay at the limits of $\mu \pm \sigma$.

In any event, the errors presented in figure 14 are based on item calibrations obtained with publisher data. IRT item calibrations based on ECLS-B data would yield somewhat different standard errors. Items comprising the 9-month BSF-R mental scale, by item set, together with IRT item difficulty and discrimination parameters obtained using publisher data are reported in table 4.

Construction of the 9-month BSF-R motor scale followed the same procedures summarized above for the mental scale. The reduced motor scale contains 35 items, including 11 basal, 14 core, and 10 ceiling items. If the core item set raw score is 0, 1, 2 or 3 points, the basal item set was administered. If the core item set raw score is 12, 13, or 14 points, the ceiling item set was administered.

Figure 14. Standard error of measurement for the 9-month BSF-R mental scale: IRT 2-PL item calibrations using BSID-II standardization sample: 1993



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

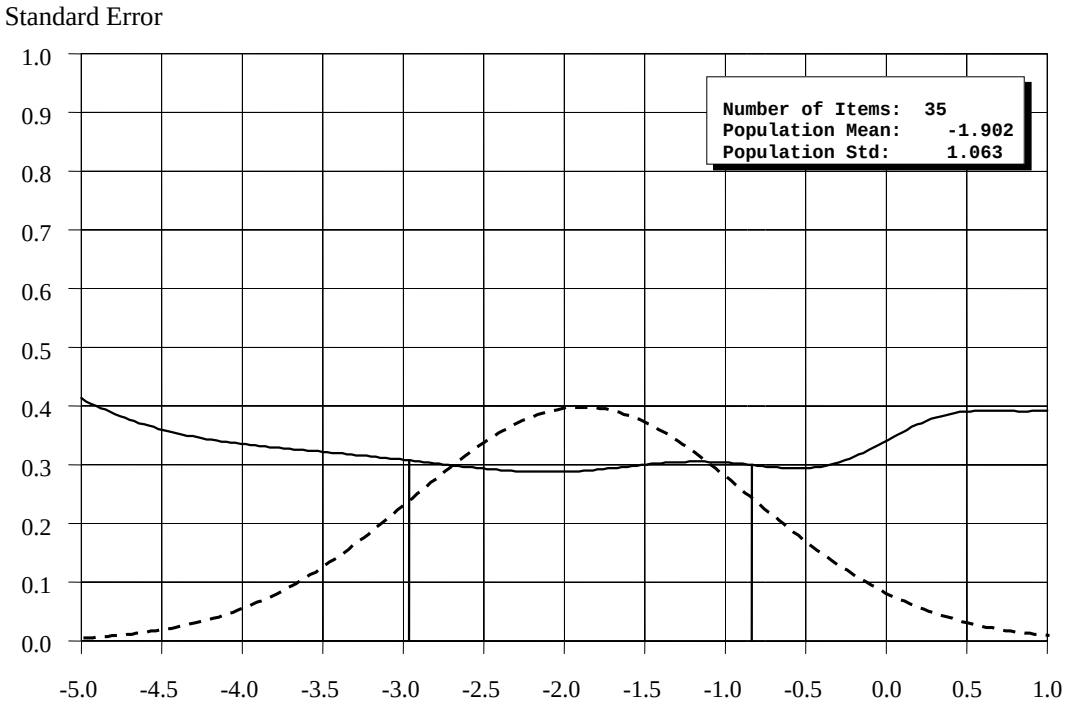
Table 4. BSID-II IRT difficulty parameter *b* and discrimination parameter *a* for items comprising the 9-month BSF-R mental scale: IRT 2-PL item calibrations using BSID-II standardization sample: 1993

Item	Item label	Item set	IRT difficulty	IRT discrimination
MEN045	Picks up cube	Basal	-4.82	2.48
MEN048	Plays with string	Basal	-4.80	1.82
MEN051	Regards pellet	Basal	-4.47	0.64
MEN052	Bangs in play	Basal	-3.93	1.15
MEN053	Reaches for second cube	Basal	-3.82	1.21
MEN055	Lifts inverted cup	Basal	-4.36	1.40
MEN057	Picks up cube deftly	Basal	-3.77	1.16
MEN058	Retains two cubes for 3 seconds	Basal	-3.18	0.83
MEN059	Manipulates bell, showing interest in detail	Core	-2.96	1.64
MEN061	Vocalizes three different vowel sounds	Basal	-3.34	0.93
MEN062	Pulls string adaptively to secure ring	Core	-2.65	1.10
MEN065	Retains two of three cubes for 3 seconds	Core	-2.40	1.62
MEN066	Rings bell purposely	Core	-2.39	1.55
MEN069	Looks at pictures in book	Core	-2.19	1.81
MEN071	Repeats vowel-consonant combinations	Core	-1.24	1.19
MEN072	Looks for contents of box	Core	-1.49	1.32
MEN073	Turns pages of book	Core	-1.18	1.52
MEN076	Jabbers expressively	Core	-0.75	0.94
MEN077	Pushes car	Core	-1.32	1.38
MEN079	Fingers holes in pegboard	Core	-1.32	1.24
MEN081	Responds to spoken request	Core	-1.01	1.23
MEN086	Puts three cubes in cup	Core	-0.52	1.50
MEN089	Puts six beads in box	Ceiling	-0.28	1.51
MEN091	Scribbles spontaneously	Ceiling	0.00	0.98
MEN095	Puts nine cubes in cup	Ceiling	0.70	0.95
MEN099	Points to two pictures	Ceiling	1.96	1.06
MEN100	Uses two different words appropriately	Ceiling	0.74	1.31
MEN101	Shows shoes, other clothing, or object	Ceiling	0.76	1.56
MEN102	Retrieves toy (visible displacements)	Ceiling	1.03	1.09
MEN104	Uses rod to attain toy	Ceiling	1.02	1.17
MEN106	Uses word(s) to make wants known	Ceiling	1.63	1.95

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

Figure 15 shows standard errors for the 9-month BSF-R motor scale based on item calibrations obtained with publisher data. IRT item calibrations based on ECLS-B data yield somewhat different standard errors. Items composing the BSF-R motor scale, by item set, together with IRT item difficulty and discrimination parameters obtained using publisher data are reported in table 5.

Figure 15. Standard error of measurement for the 9-month BSF-R motor scale: IRT 2-PL item calibrations using BSID-II standardization sample: 1993



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

Table 5. BSID-II IRT difficulty parameter *b* and discrimination parameter *a* for items of the 9-month BSF-R motor scale: IRT 2-PL item calibrations using BSID-II standardization sample: 1993

Item	Item label	Item set	IRT difficulty	IRT discrimination
MOT022	Sits with slight support for 10 seconds	Basal	-4.62	1.10
MOT025	Shifts weight on arms	Basal	-5.09	1.10
MOT026	Turns from back to side	Basal	-4.46	0.93
MOT028	Sits alone momentarily	Basal	-4.27	1.09
MOT029	Uses whole hand to grasp rod	Basal	-4.98	1.04
MOT030	Reaches unilaterally	Basal	-4.43	0.85
MOT031	Uses partial thumb opposition to grasp cube	Basal	-3.95	1.20
MOT032	Attempts to secure pellet	Basal	-3.84	1.08
MOT034	Sits alone for 30 seconds	Basal	-3.27	1.03
MOT036	Sits alone steadily	Basal	-3.34	0.93
MOT038	Turns from back to stomach	Basal	-3.88	1.05
MOT040	Makes early stepping movements	Core	-3.02	1.40
MOT041	Uses whole hand to grasp pellet	Core	-2.92	0.88
MOT043	Moves forward using prewalking methods	Core	-2.94	0.99
MOT044	Supports weight momentarily	Core	-2.80	0.60
MOT045	Pulls to standing position	Core	-2.53	1.18
MOT046	Shifts weight while standing	Core	-2.29	1.36
MOT049	Uses partial thumb opposition to grasp pellet	Core	-2.68	0.70
MOT051	Moves from sitting to creeping position	Core	-2.43	1.48
MOT052	Raises self to standing position	Core	-2.02	1.86
MOT053	Attempts to walk	Core	-1.74	1.20
MOT054	Walks sideways while holding on to furniture	Core	-1.62	1.76
MOT058	Grasps pencil at farthest end	Core	-0.76	1.39
MOT059	Stands up I	Ceiling	-0.49	1.39
MOT060	Walks with help	Core	-1.19	1.65
MOT061	Stands alone	Core	-0.67	1.72
MOT062	Walks alone	Ceiling	-0.34	2.27
MOT063	Walks alone with good coordination	Ceiling	-0.30	0.92
MOT065	Squats briefly	Ceiling	1.37	1.10
MOT067	Walks backward	Ceiling	0.75	1.11
MOT068	Stands up II	Ceiling	0.98	1.09
MOT070	Grasps pencil at middle	Ceiling	0.67	0.57
MOT071	Walks sideways	Ceiling	1.00	0.85
MOT072	Stands on right foot with help	Ceiling	0.99	1.29
MOT073	Stands on left foot with help	Ceiling	1.19	1.40

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

Using IRT analyses identical to those used for the BSF-R mental scale, the 5-month age item set was identified as the best source for basal items and the 13-month age item set was identified as the best source for ceiling items. Although all the basal items were obtained from the 5-month age item set, they ranged in age from 3 to 7 months. Although all the ceiling items were obtained from the 13-month age item set, they ranged in age from 8 to 22 months. In addition, as with the BSF-R mental scale, IRT analyses were used to define the rules for applying the basal or ceiling items for the motor scale. IRT analyses determined that the basal items should be applied if the child received credit for 3 or fewer core motor items, which is the equivalent of 1 standard deviation below the mean. If the child received credit for 11 or more motor items in the core set, then the ceiling items would be applied.

2.1.10 Reformulation of Administration Booklet, Training Materials, and Video

On the basis of publisher data, the ECLS-B's BSF-R should successfully measure ability well out into the tails of the distribution, with standard errors of less than 0.4 and a forecast reliability of 0.80 or greater. This assumes, of course, that field interviewers did as well in the home as trained clinicians and researchers do in clinical and research settings. Training the ECLS-B field interviewers to the high standard that would obtain strong results required further modifications to the BSF-R administration materials and to the training approach.

During the fall 1999 field test, ECLS-B interviewers complained that the administration booklet and scoring sheets were confusing. Furthermore, results of the fall 1999 field test showed that interviewers did not always understand when it was necessary to administer basal item sets or ceiling item sets. ECLS-B interviewers also found the flexibility of BSID-II administration to be challenging. A trained clinician will move fluidly among the items, reordering items freely based on the child's interest and ability and readministering those that the child was not interested in when first presented. However, ECLS-B interviewers needed more explicit structure than do clinicians and researchers adept at administering the BSID-II.

As a result, the administration of the BSF-R was made more straightforward by simplifying step-by-step administration instructions and by providing more structure. 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

closely corresponded with the presentation of these items in the Bayley manual. This being said, efforts were taken to streamline the administration and to standardize scoring for field staff.

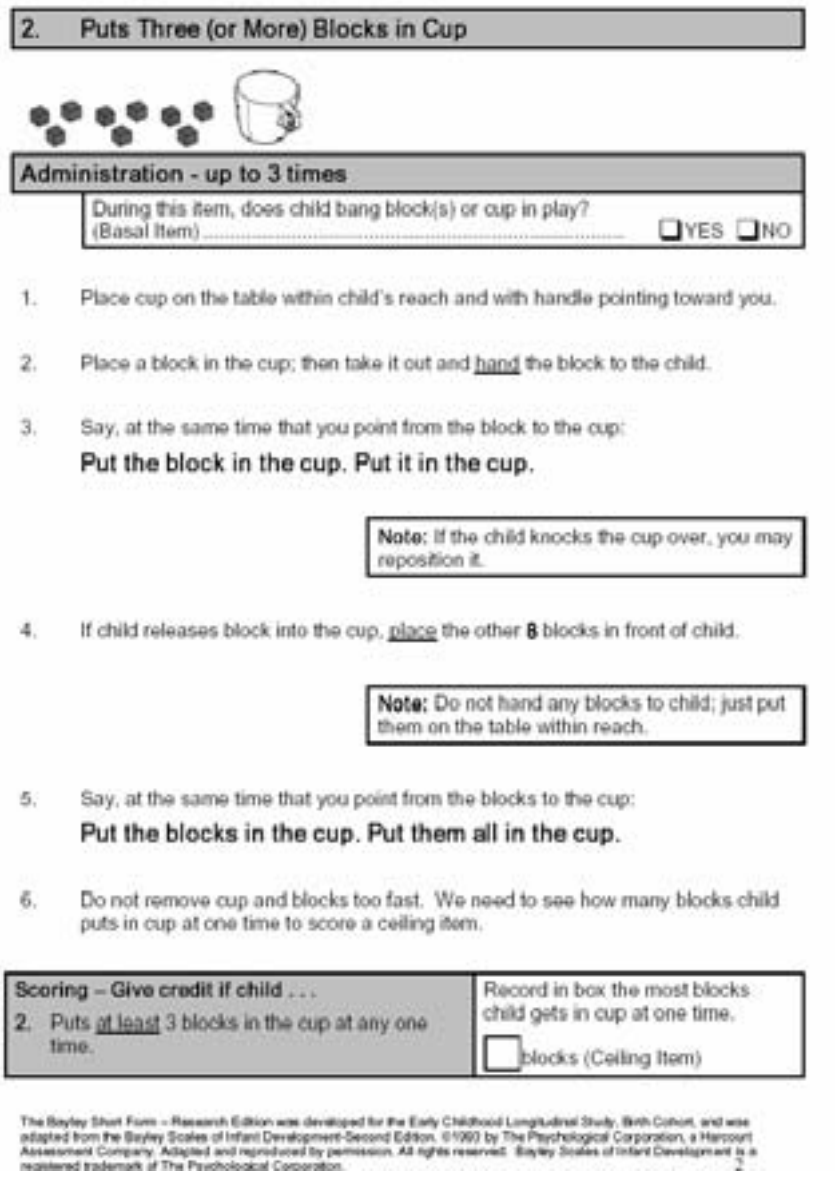
A more structured layout was created for each item. First, item number and item name were located at the top of the page, with a picture of the materials used right underneath, e.g., cubes and cup (exhibit 3). The next bar is Administration, with the number of permissible administrations highlighted, e.g., up to 3 times, and administration instructions right underneath, in this example steps 1.6.

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 was to allow the child ample opportunity to put as many as 9 blocks in the cup so that a ceiling item could be scored from the administration of this core item.

Note boxes were also included, like those over on the right side of the page, 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 could arise. The scoring criteria were highlighted in the box at the bottom, and special instructions were included to cover any special situations, such as, if the child walks with good balance and coordination, then the interviewer should also give credit for “Walks independently” and “Walks with help.”

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

Exhibit 3. Sample BSF-R administration page, 9-month data collection: 2001–02



2. Puts Three (or More) Blocks in Cup

Administration - up to 3 times

During this item, does child bang block(s) or cup in play? (Basal Item) ..... ❑YES ❑NO

1. Place cup on the table within child's reach and with handle pointing toward you.
2. Place a block in the cup; then take it out and hand the block to the child.
3. Say, at the same time that you point from the block to the cup:
   Put the block in the cup. Put it in the cup.

Note: If the child knocks the cup over, you may reposition it.

4. If child releases block into the cup, place the other **8** blocks in front of child.

Note: Do not hand any blocks to child; just put them on the table within reach.

5. Say, at the same time that you point from the blocks to the cup:
   Put the blocks in the cup. Put them all in the cup.
6. Do not remove cup and blocks too fast. We need to see how many blocks child puts in cup at one time to score a ceiling item.

Scoring – Give credit if child . . .	Record in box the most blocks child gets in cup at one time.
2. Puts at least 3 blocks in the cup at any one time.	☐ blocks (Ceiling Item)

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), Nine-Month Data Collection, 2001–02.

The application of the basal and ceiling rules was also simplified. 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 Cs (credits) and NCs (no credits) the child receives and to continue administering additional age sets until the criterion has been satisfied. Given the limited background of the ECLS-B interviewers and the demands of the ECLS-B home visit, the basal-ceiling rule needed to be simplified. The goal was a

single set of basal items and a single set of ceiling items with clear rules for their administration. Therefore, the basal item set had an extended age range, from 4 months to 8 months on the mental scale and from 3 months to 8 months on the motor scale. The ceiling set ranged from 11 months to 19 months on the mental scale and from 8 months to 22 months on the motor scale. Because the BSF-R is a reduced set of items, the rules to determine whether or not to proceed to a basal set of items or a ceiling set of items needed to be revised proportionately. The basal and ceiling rules were established using IRT analyses. The basal and ceiling rules were also clarified to help interviewers determine whether or not they needed to do these supplementary items, by making the instructions as specific as possible. On the mental scale, if there were only 0 to 3 "Cs," then the interviewer should administer a single basal item (the rest of the basal items being scored by observation); if there are 9 or more "Cs," then the interviewer should administer the set of ceiling items. On the motor scale, if there are only 0 to 3 "Cs," then administer the basal items, and if there are 12 to 14 "Cs," then administer the ceiling items.

2.1.11 Chapter Summary

This chapter described the work that was done to develop the Bayley Short Form-Research Edition (BSF-R), a shortened form of the Bayley Scales of Infant Development, Second Edition (BSID-II). The first step was to obtain the standardization dataset of the BSID-II from the publisher, the Psychological Corporation. The second step was to conduct Item Response Theory (IRT) analysis of the all the BSID-II items in the standardization dataset. This step produced ability parameter values for each item as well as discrimination parameter values. The items were then essentially "lined up" in their order of increasing ability (or difficulty). Then items in the publisher's 8- to 10-month mental and motor item age sets were identified and singled out for further examination of their discrimination parameters. The discrimination parameter represents an item's discrimination power--that is, how well the item differentiates between individuals at higher and lower abilities (scale scores). An item that can be responded to correctly by either group is a poor discriminator. Items with good discrimination parameter values, i.e., 0.7 or greater, were identified a candidates for the BSF-R. Additional item selection criteria were also applied, further eliminating items from the list of candidate items, including simplicity of administration, objectivity of scoring, and minimization of necessary materials.

The best items in the 8- to 10-month age sets were selected for the core item sets of the mental and motor scales. These items would be administered to all children in the ECLS-B. However, supplementary item sets, similar to a routing task, would be necessary for children in the tails of the ability distribution, that is children at ability levels that were 1 or more standard deviations from the mean

of the target ability distribution. Therefore, items with ability parameter values that were 1 standard deviation or greater from the mean of the core set items were identified for the set of basal items and for the set of ceiling items. Operational criteria were also taken into consideration when selecting the basal and ceiling items. It was expected that 16 percent of children would require administration of the basal items, and 16 percent would require the ceiling items, in addition to administration of the core items. (Under no circumstances would a child be administered both the basal set and the ceiling set for the BSF-R mental scale or for the BSF-R motor scale.)

For the mental scale, a set of 13 items from the target age set with good discrimination parameter values formed the core set of items that would be administered to all children in the ECLS-B. A set of 9 items from the 5-month item age set were identified as good items for the basal set of items. A set of 9 items from the 13-month item age set were identified as good items for the ceiling set of items. Basal and ceiling rules were established on the basis of IRT analysis to route children to either the basal item set or to the ceiling item set, as necessary.

For the motor scale, a set of 14 items from the target age set with good discrimination parameter values formed the core set of items that would be administered to all children in the ECLS-B. A set of 8 items from the 5-month item age set were identified as good items for the basal set of items. These were supplemented by two items from the 4-month item age set because these two items could be scored from another basal item already included, thereby offering even greater coverage at the lowest end of the ability distribution. A set of 10 items from the 12-month item age set were identified as good items for the ceiling set of items. Basal and ceiling rules were established on the basis of IRT analysis to route children to either the basal item set or to the ceiling item set, as necessary.

This page is intentionally left blank.

3. TESTING THE BAYLEY SHORT FORM–RESEARCH EDITION IN THE FALL 2000 FIELD TEST

The first step was the selection of items and determination of the appropriate structure and rules for the basal and ceiling item sets, described in chapter 2. The next step was to ensure that the Bayley Short Form–Research Edition (BSF-R) would obtain reliable scores when administered by field staff in a home setting. This required testing and revision to ensure the psychometric strength of the BSF-R. The following sections summarize the field testing procedures that were undertaken to identify items that were weak psychometrically and to tailor training procedures to ensure that inexperienced field staff administered and scored the BSF-R items reliably. In addition, the procedures for production of a standardized BSF-R training video are presented.

3.1 Results From the Field

The fall 2000 9-month field test gave the first indication of how well the 9-month BSF-R performed. First, the ECLS-B interviewers were relieved to see a reduction in the number of items for the BSF-R, the streamlined procedures, and the formatting of the administration booklet. From an operational point of view, the finding that the BSF-R was taking much less time to complete was cause for optimism. The average BSF-R administration time for the fall 2000 field test was under 26 minutes, down from 43 minutes for the full Bayley Scales of Infant Development, Second Edition (BSID-II) during the fall 1999 field test. In addition, the BSF-R completion rate improved: of 722 cases with a completed Parent Interview, 706 had a completed BSF-R, for a BSF-R completion rate of 97 percent. This was an improvement from the fall 1999 field , where only 77 percent of cases with a completed parent interview had a completed BSF-R.

The real measure of success, of course, was psychometric. Fall 2000 9-month field test data were equated with the publisher data, using item response theory (IRT) techniques. If the equating proved successful, then the BSF-R scores could be put on the same scale metric as publisher data, and then analysts could make use of the published mental scale and motor scale index scores. This would make the scores obtained in ECLS-B comparable with other studies that have used the full BSID-II. In addition, IRT true score equating helps to identify problematic items that are misbehaving. For example, if an item with a difficulty parameter value of 2.0 (based on publisher standardization data) has a difficulty

parameter value of 1.0 on the basis of ECLS-B data, then it can be concluded that the ECLS-B interviewers were too lenient in their scoring of this item and that too many children were given credit for it when they probably should not have been given credit. This provides the opportunity to determine whether interviewers were incorrectly administering or incorrectly scoring items in a systematic manner.

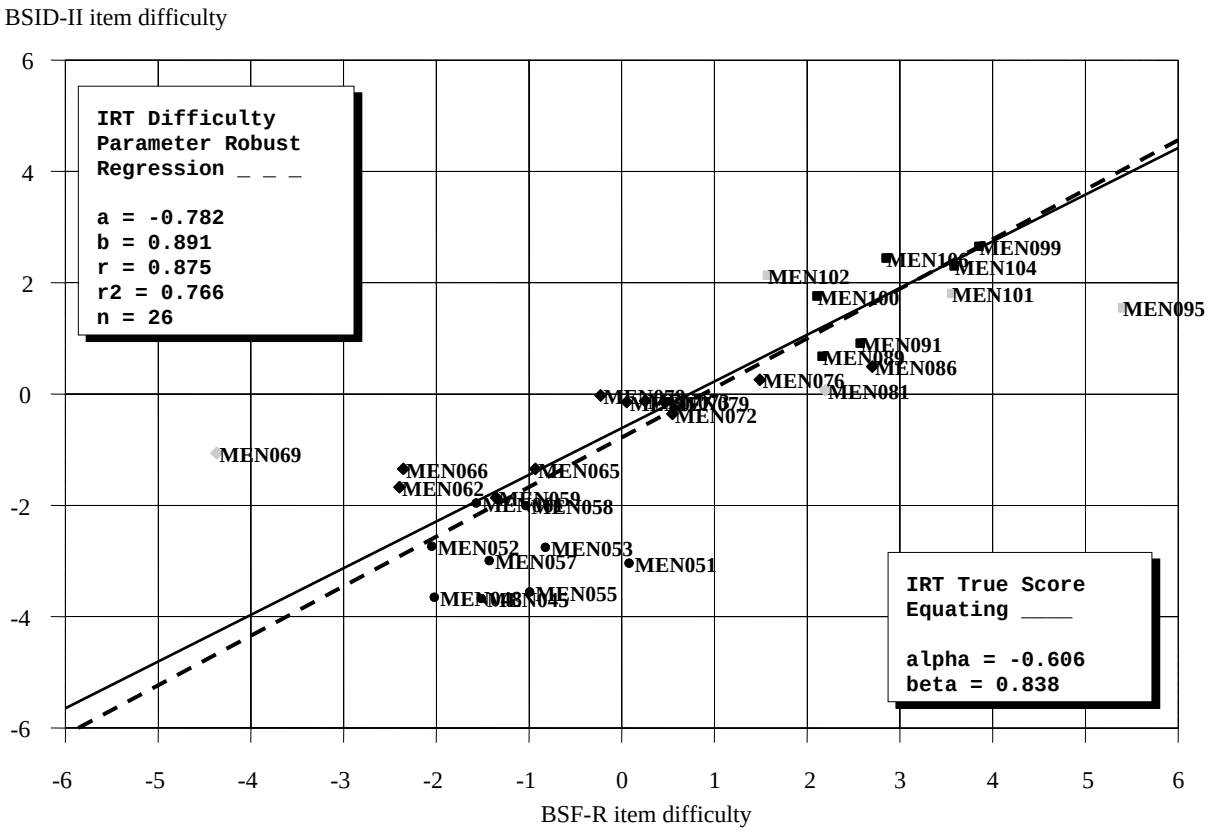
3.2 Identification of Nine-Month BSF-R Problem Items

Findings from the true score equating of the field test data were quite helpful for training purposes. IRT equating identified items that were problems (i.e., they showed up as "outliers" on the graphs). The content of the item and review of item scoring results were reviewed to identify any scoring bias. The problem items were identified and then examined in order to target the causes of the problems. Once the nature of the problem with an item was understood, then strategies were devised both to remedy the problem and to fine-tune the training on the item to help interviewers understand the items better. For example, interviewers seemed to have trouble scoring the "pencil items" (Holds pencil at farthest end, Holds pencil at middle). Review of the objective of the item and consultation with Dr. Matula of the child assessment work group suggested that interviewers may not have understood that these items are intended to assess children's pre-writing skills. Therefore, the training was revised to emphasize that these items assess pre-writing and that children show a developmental progression in which the child first holds the pencil in a whole hand grasp at the end farthest from the point, later making the transition to holding the pencil at the middle, and, later, at the end nearest the point.

Figure 16 shows how well the 9-month BSF-R mental scale equated with publisher data. Ideally, the items should distribute themselves around a straight line. If they do, then a linear transformation of origin and scale to transform the BSF-R scores into BSID-II standardized index scores can be done.

The dashed line shows the linear relationship between the field test and publisher item difficulties. The solid line takes into consideration both the item difficulty and discrimination parameters. The $r^2 = .74$ shows that 74 percent of the variance in item difficulty parameters is common to both sets of calibrations. Several items are easily identified as outliers. Three items to the left of the figure, MEN062, MEN066 and MEN069, are core items that performed much like basal items on the field test (i.e., nearly all of the children in the field test received credit for these items). To the right of the figure, MEN095 proved to be much more difficult on the field test than it was in publisher data. However, the majority of the items seem to be working fairly well.

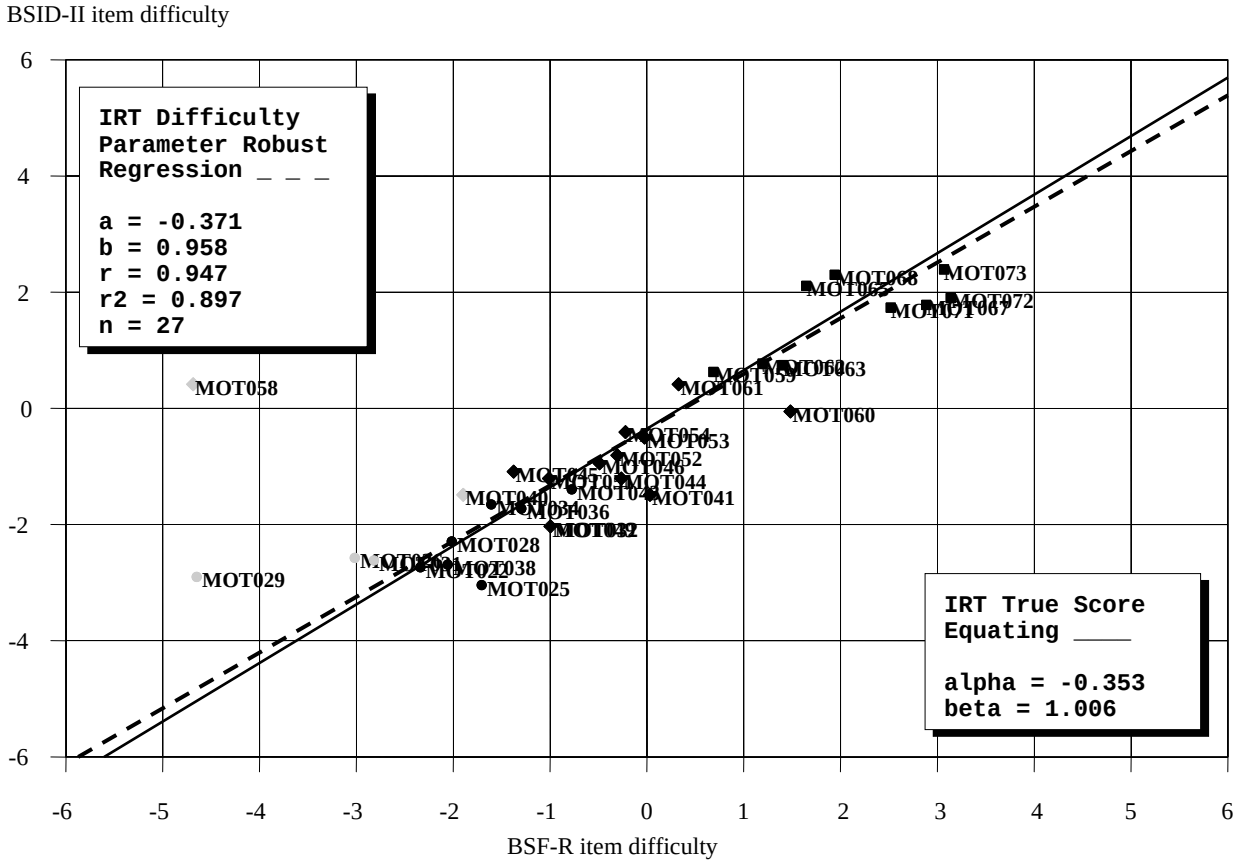
Figure 16. BSF-R 9-month mental scale equating of field test data and BSID-II standardization sample: 1993 and fall 2000



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Fall 2000 Field Test and Psychological Corporation publisher data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), 1993.

The alignment of BSF-R motor scale items (figure 17) was noticeably better, as reflected in an r^2 of .88. Again, a few items diverged considerably from the publisher standard: MOT058 was found far to the left of the regression line and MOT070 far to the right (although this item is hard to identify in this graph). Both of these are pencil items—"Holds pencil at nearest end" and "Holds pencil at farthest end." Apparently, interviewers misinterpreted the concept of holding a pencil at farthest (and nearest) end and were giving credit if the child merely picked up the pencil. They were not judging whether the child showed any intent to write first and THEN held the pencil at the farthest (or nearest) end. This led to a major change both to the administration booklet (a checkbox was added: Did child show intent to write?) and to our training approach in which we focused on recognizing intent to write first and then noticing where the child held the pencil.

Figure 17. BSF-R 9-month motor scale equating of field test data and BSID-II standardization sample: 1993 and fall 2000



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Fall 2000 Field Test and Psychological Corporation publisher data set for the Bayley Scales of Infant Development,Second Edition (BSID-II), 1993.

3.3 Training Procedures and Certification on the Bayley Short Form–Research Edition

On the basis of the results of the IRT analyses and benefiting from the comments of interviewers during debriefing, a number of steps were taken to revise the BSF-R administration booklet and to revise and improve the training approach. These steps are described in the following sections.

A series of quality control measures were instituted to make sure that interviewers were reliable administrators of the BSF-R. In order to ensure that trainees attained an acceptable level of proficiency administering the BSF-R by the end of training, a series of quality checks was implemented throughout the training.

At the first level, during an in-class session, trainees viewed a videotape of a BSF-R administration and then practiced evaluating the quality of the administration and gave scores for the child's response on each item. This gave trainees experience with the evaluation process that would be used to evaluate their administration and scoring of the BSF-R. In the next session, trainees again viewed a videotaped BSF-R administration, evaluated the accuracy of the administration and assigned item scores. For this purpose, a standard form was used, the BSF-R review form, in which administration procedures were listed for each item and the trainee recorded whether or not each step was administered correctly. Trainees also scored the child's performance for each item. Trainees had to meet a predetermined criterion of 85 percent or greater to be considered "passing." (This criterion was chosen first because it is the same criterion as that used by Mathematica Policy Research [MPR] in the Early Head Start [EHS] Research and Evaluation Project and also because 85 percent was used as the criterion on other assessments in the ECLS-B, for example, Nursing Child Assessment Teaching Scale [NCATS] coding.) Those trainees who did not reach this criterion were required to attend help labs to work on those procedures and scores that caused them problems. In order to advance to the "live practice," trainees were required to pass the videotape review at 85 percent or greater for the combination of administration and scoring accuracy.

At the live practice session, trainees were videotaped while they administered the BSF-R. All videotapes were reviewed before the end of training by Westat child development staff members who had expertise in the BSID-II and had participated in the training. Any trainee who scored between 80 and 85 percent on the BSF-R administration was given a second chance to administer the BSF-R to another child. If at that time the trainee scored 80 percent or better but not 85 percent, that trainee was followed closely after training and was not allowed to begin gathering data until a child development staff member had arrived in the field and conducted further BSF-R training. After the individual had received further training and practice, the child development staff member reviewed the individual's administration in the field and completed a BSF-R review form. By this stage in the review process, all trainees passed their BSF-R administration at 85 percent accuracy, which was the important criterion.

Finally, during the field test data collection, it was planned that a videotape of a BSF-R administration by a Westat child development staff member would be sent out to obtain reliability data. Field staff were expected to view the videotape and rate both the accuracy of item administration and to score the child's performance on each item using the same quality control form that was used at training to score their live practice performance. Interviewers were then expected to return this quality control form to Westat for scoring to ensure reliability of BSF-R administration and item scoring. The purpose of this step in the field test was to assess quality control methods and the ability of home office staff to obtain the videotapes, duplicate sufficient copies, and design and review the measurement of interviewer reliability in a timely manner. Because the procedure and scoring criteria were revised repeatedly during the field test, there was no fair way to establish reliability across all interviewers or summarize the results from the field test, however.

For the national study, a single quality control videotape was considered insufficient to prevent interviewers from drifting from the standard over the course of an entire year of data collection and a decision was made to increase this to four videotaped BSF-R administrations to be sent to interviewers quarterly over the year-long data collection period to obtain reliability data for the national study in 2001. For further information about BSF-R quality control reviews, please see section 4.2, below, and the *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Design and Operations Report for the Nine-Month Data Collection* (U.S. Department of Education, National Center for Education Statistics, 2004 [unpublished report]).

3.4 Further Revisions to Bayley Short Form–Research Edition Administration Booklet

IRT analyses identified several items as problematic. In some cases, careful review of the BSF-R administration instructions in the booklet found ambiguities or lack of specificity in either the administration steps or the scoring instructions. For those items in which the lack of specificity originated in the BSID-II manual, Westat child development staff consulted with the Psychological Corporation staff member who had been instrumental in the 1993 restandardization of the BSID. Her clarifications were incorporated into the administration booklet and into the training scripts. For example, "Puts 3 blocks in cup" can be administered three times. However it is not clear what constitutes a single administration for this item because it combines "Puts 1 block in cup," and "Puts 9 blocks in cup" does each item constitute an administration? For other items, simple instructions were added, reminding interviewers to observe specific aspects of the child's response to improve the accuracy of scoring. For example, the following

instruction was added to the administration instructions for the item "Pulls string purposely to grab ring" in order to focus interviewers' attention to the child's attempts to bring the ring close enough to grab: "Watch what child does to bring the ring close enough to grab it. WATCH whether child grabs the ring." In addition, the titles of a few items were edited, as minimally as possible, to make the purpose of the item more salient. For instance, "Pulls string purposely to grab ring" was changed from the original "Pulls string adaptively to secure ring" to highlight the importance of actually grabbing the ring in order to give credit for the item as well as to highlight the intentionality of the string pulling.

Formatting changes in the administration booklet also made it visually easier to follow the directions. These were changes to the appearance of the pages and layout of the scoring sheets and not to material referring to the content of the items.

3.5 Standardized Training Video

The national training was expected to have approximately 240 trainees in approximately 13 training rooms. Westat has five child development staff members who are familiar with the BSID-II. Ensuring equality of training on the BSF-R across all the rooms would be a challenge. To this end, an expanded BSF-R training video was developed to present the BSF-R information to all trainees in a standardized way. The content of the video was also expanded from previous field tests to include the purpose of the item, highlights of what to look for in the child's response, what a "C" (credit) performance and a "NC" (no credit) performance look like, as well as solutions for often-encountered troublesome situations. The expanded video for the national training also included basic information about how to set up the testing situation and how to apply the basal and ceiling rules. In short, the video was a self-contained training on the BSF-R that interviewer trainees could take with them at the session's end. It was intended that interviewers could review the video at home as necessary to learn or refresh their knowledge about the administration and scoring of the BSF-R.

3.6 Chapter Summary

IRT analysis of the fall 2000 9-month BSF-R data identified several items that did not yield the same item parameters as in the BSID-II standardization sample. On all but one of these items, ECLS-B interviewers tended to give credit to children too easily. That is, the ability parameter of the BSF-R items located them lower in the ability distribution compared with the ability parameter value of the same

items in the BSID-II standardization sample. In essence, the items behaved as if they were basal items. Therefore, these items showed up as outliers on the equating graphs. Only one item on the mental scale turned out to have a higher ability parameter value than on the field test.

On the BSF-R mental scale, the items that behaved like basal items included: MEN062 (Pulls string adaptively to secure ring), MEN066 (Rings bell purposely), MEN069 (Looks at pictures in book). MEN095 (Puts nine cubes in cup) was the only BSF-R mental item that was more difficult on the field test than in the standardization sample.

On the BSF-R motor scale, only items MOT058 (Grasps pencil at farthest end) and MOT070 (Grasps pencil at middle) were scored as easier by field test interviewers than their counterparts in the BSID-II standardization sample. No motor items were identified as more difficult on the field test than in the standardization sample.

After the problem items were identified, they were reviewed for misinterpretation, misrepresentation of the administration materials, and misrepresentation of the scoring instructions. Revisions were made to field materials, training instructions and to the training videotape, as appropriate.

4. NINE-MONTH NATIONAL TRAINING AND FIELD RESULTS

The ECLS-B followed rigorous quality control procedures to ensure the quality of the Bayley Short Form–Research Edition (BSF-R) data. These procedures included standardization of training, high certification standards during live practice sessions at training, and ongoing quality control of BSF-R administration during the year of data collection including quarterly reviews of BSF-R administration and scoring with rigorous monitoring and intervention for any field staff not meeting these standards. In addition, psychometric analyses were undertaken to ensure that data collected were consistent with the publisher's standardization data set. These procedures are presented in the following section.

4.1 Training and Certification Results

For the national training, certification procedures were more rigorous than for the fall 2000 field test. Certification on the BSF-R involved three aspects. Trainees were evaluated on their ability to administer the items according to the standardized instructions in the administration booklet, their knowledge of each item's scoring criteria, and their ability to interpret children's responses. The certification process included three steps, the first two of which served to identify those needing further training and the last one of which was considered the criterion reference for "passing the BSF-R." On the first step, trainees (in class) reviewed a videotaped BSF-R administration and scored the examiner's accuracy of administration and assigned their own scores to the child's responses on each item. Essentially, they were reviewing the examiner's administration in order to identify any errors in administration, and they viewed the child's responses in order to practice scoring the items. This procedure served as a "practice quiz" that prepared them for the real quiz to follow. It also exposed them to the review process by which they would be judged at the "live practice session" during training and during field quality control visits.

The second step was a similar video review that followed in the next session, which served as a screener to identify individuals having problems understanding the administration rules and the scoring criteria. For this quiz, 85 percent was considered passing for accuracy of administration, and 90 percent was considered passing for accuracy of scoring. Any trainee who did not pass at 85 percent or higher on administration or 90 percent or higher on scoring accuracy was required to attend a help lab to work on administration skills and/or understanding the scoring criteria. The requirement to score 90

percent or higher on scoring accuracy was imposed because scoring errors can result in misapplication of the basal/ceiling rule with resulting loss of data.

The final step in the certification process was live practice. At live practice, each trainee was required to complete the BSF-R with a child recruited for the training. Children recruited ranged in age from about 8 months to 12 months, the approximate age range trainees would be encountering in the field. Trainees were paired up so that one administered the BSF-R while the other videotaped the administration and child's responses. Trainees then switched roles for a second child "volunteer." Each videotape was reviewed by the trainees' lead trainers before the end of training. Each videotape was reviewed for accuracy of administration and for accuracy of scoring using the same standardized BSF-R review form that had been used during previous video practice sessions. In order to become certified to administer the BSF-R, each trainee had to score 85 percent or higher for accuracy of administration and 90 percent or higher for accuracy of scoring. Trainees who had not quite passed but showed potential to improve were given the option of doing a second live practice BSF-R. Those who did not achieve certification levels and showed no potential for improvement were released from the project. Two training sessions were held, the first one in October 2001 with 202 trainees, and the second (attrition) training in December 2001 with 41 trainees. A total of 14 trainees were released as a result of poor BSF-R administration from the two trainings. In addition, two trainees resigned from the attrition training rather than repeat a live practice session. These two individuals showed potential to learn the BSF-R, although they chose not to continue with the study. Pooling the scores of trainees on the BSF-R across the two trainings resulted in an average BSF-R administration score of 93 percent and 92 percent for scoring accuracy.

4.2 Results From the Field

The BSF-R completion rate (for the BSF-R, completion refers to the complete administration of at least one scale, either the mental scale or the motor scale) is 99 percent, calculated as the number of completed BSF-R administrations divided by the number of completed parent interviews. For further information about completion rates, please refer to *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005-013). When interviewers administered the BSF-R, they recorded the start time of the BSF-R on the first page of the administration pages and recorded the end time when they had completed all BSF-R items and had asked the primary caregiver the two questions about how typical the child's performance

was. Time to complete the BSF-R averaged approximately 36 minutes, which is longer than in the field test. This is probably due to more variability in the performance of the larger number of interviewers and because of the greater variability in the sampled children (e.g., the national sample contained an oversample of very low birth weight children.)

Quality control visits were completed for all interviewers continuing with the study. For most interviewers, the quality control visit was done by the field supervisor. Average scores for BSF-R administration for these interviewers were 96 percent for administration and 85 percent for scoring.

During the entire year of 9-month data collection, four quality control videotapes were sent to interviewers on a quarterly basis in order to prevent interviewer drift from BSF-R standards. The interviewer was instructed to view the videotape twice. On one pass through the tape, the interviewer was instructed to evaluate the accuracy of the assessor's administration of the items using the same standardized BSF-R quality control form that was used during training. This included the assessor's correct implementation of the basal and ceiling rules. On the other pass through the videotape, the interviewer was instructed to score the child's performance for each item that was administered as if the interviewer had administered the BSF-R to the child. The order in which the interviewer completed these passes through the videotape (i.e., assessor accuracy first and item scoring second, or vice versa) was left to the interviewer's discretion.

The total average for the four quality control videotapes for administration was 95.3 percent and 95.0 percent for scoring. If an interviewer did not pass a quality control videotape, the interviewer received direct feedback via telephone from one of the Westat child development staff members who reviewed the errors and determined whether the interviewer was making similar errors in the field. If an interviewer did not pass two quality control videotapes in a row, the interviewer's supervisor was contacted, and the interviewer was required to videotape her or his next BSF-R administration in the field for review by a Westat child development staff member. The videotape was express mailed back to Westat and was evaluated within a day or two of receipt, resulting in no work loss for the interviewer. Eight interviewers were required to submit videotapes of a BSF-R administration for quality control review by Westat child development staff. Six did quite well and passed this quality control review on the first try. One interviewer was given two days of additional BSF-R training in the field by her field supervisor who then conducted a quality control review of the interviewer's BSF-R administration in the field. This interviewer then passed that BSF-R review. Only one interviewer was released from the study due to an inability to adhere to the quality control standards. This interviewer had completed relatively

few cases; this interviewer's data were reviewed, and item response theory (IRT) analyses did not identify the BSF-R scores as problematic.

4.3 Results of Post-Field Test Item Modifications

After the fall 2000 field test, several improvements were made to BSF-R training and to the BSF-R administration booklet in an attempt to correct identified problem items. True score equating of the 9-month national data showed that different problems emerged for the 9-month mental scale and for the motor scale. Several mental items that had been a problem during the field test were no longer identified as problems. Instead, several new items (MEN045, MEN055, and MEN062) were identified as problematic and were ultimately dropped from the IRT equating of the BSF-R. These items continued to be administered in the field and these item scores were included in the total scores. Several motor items that had been a problem during the field test were no longer identified as problems. Instead, several new items were identified as problematic, including MOT025, MOT028, MOT030, MOT034, MOT046, MOT052, and MOT062. These items were excluded from the IRT equating of the BSF-R motor scale but continued to be administered in the field and were included in the total scores. For further discussion about this issue, see section 4.4, which follows.

4.4 Item Response Theory Item Calibrations of Nine-Month National ECLS-B Data

For the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), the Bayley Scales of Infant Development, Second Edition (BSID-II) have been adapted for use in household survey interviews. Sets of basal, core, and ceiling items for the BSF-R were selected using item calibrations based on the publisher's standardization data set. The technical qualities of the 9-month BSF-R were then assessed using ECLS-B data. The following analyses are based on the full 9-month data collection and are based on the child weight (W1CO). The objective is to determine whether the short form scales, specially developed for use as part of a household survey interview, provide reliable measures of mental and motor development comparable to those obtained with the BSID-II.

To select items for the BSF-R, the complete sets of 178 mental and 111 motor test items in BSID-II were initially calibrated with an IRT 2-parameter logistic (2-PL) model using publisher data. The BSF-R scales have since been calibrated with an identical IRT model using a sample of 2,090 infants

assessed in the ECLS-B 9-month sample. This is a random sample that includes infants in the 8- to 16-month age range and resembles the publisher standardization data set, i.e., the age groups in the ECLS-B sample have equal *n*s as the publisher data set and 34 percent were 9 months old. Table 6 summarizes the age distribution of the publisher standard data set in comparison with the ECLS-B age distribution There is considerable interest in learning whether there is evidence of item parameter invariance properties, when clinical measures obtained with BSID-II are adapted for short form application as part of a household survey interview.

Table 6. Age frequency distribution of ECLS-B 9-month data and BSID-II standardization sample: 1993 and 2001–02

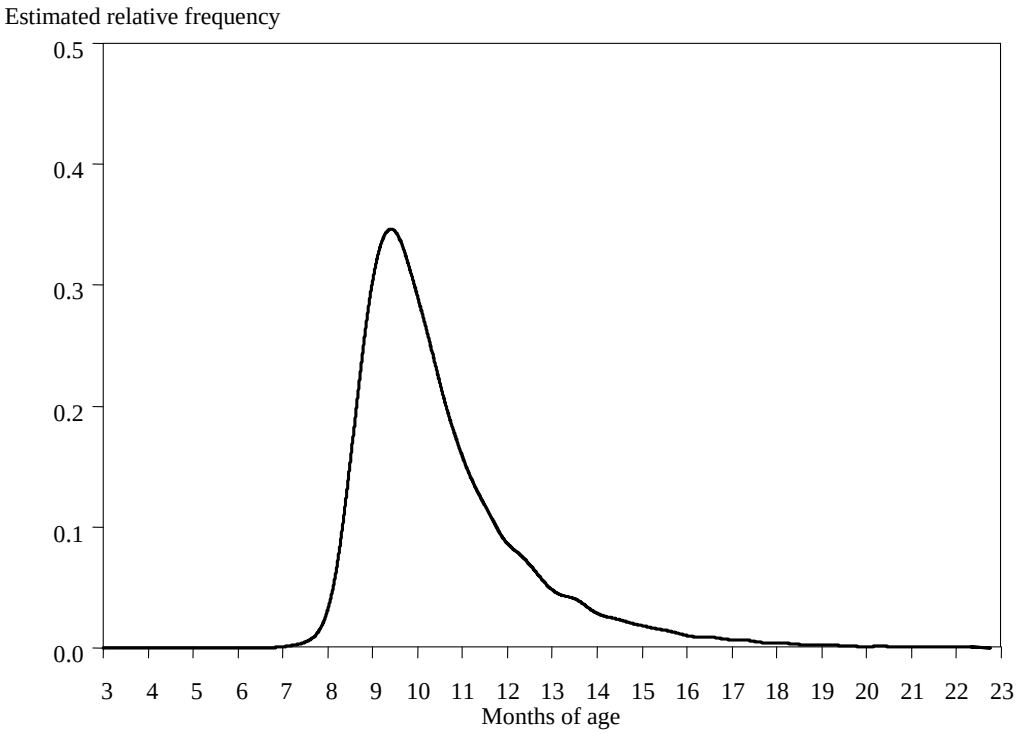
Months of age	ECLS-B 9-month data collection		ECLS-B IRT calibration sample	
	n	Percent	n	Percent
Total	10,213	100.0	2,090	100.0
3	1	0.0	†	†
6	2	0.0	†	†
7	56	0.5	†	†
8	1,557	15.2	296	14.2
9	3,501	34.3	238	11.4
10	2,235	21.9	230	11.1
11	1,166	11.4	228	11.0
12	706	6.9	238	11.4
13	391	3.8	226	10.9
14	240	2.3	219	10.5
15	149	1.5	162	7.8
16 or more	209	2.1	244	11.7

† Not applicable.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine Month Data Collection, 2001–02; standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

By way of introduction, it is almost impossible to collect a sample of infants who are exactly 9 months old. Scheduling household interviews to coincide precisely with 9 months of age would be impractical. Consequently, the 9-month target population is represented in the ECLS-B by a sample that was only approximately this age. The age frequency distribution of the 9-month national sample is reported in table 6, where frequencies are reported based on completed months of age. Indeed, a plurality of 34 percent of the infants in the ECLS-B sample have completed nine months of age, with relatively few observations (15.7 percent) below that age. The age distribution is right skew, with appreciable

numbers of 10- (21.9 percent) and 11-month (11.4 percent) infants. Most of the observations (93.6 percent) refer to infants between 8 and 13 months of age. Date of birth and date of visit are available for all of the 10,213 observations reporting information for the BSF-R. A kernel age density estimation, based on decimal months of age, is shown in figure 18. This provides a visual representation of the distribution, albeit with some smoothing of the data. (This smoothing makes it appear as if there are virtually no children beyond about 16 months of age, however.) The decimal age distribution has a mode of approximately 9.5, a median of 9.9, and an arithmetic mean of about 10.5 months of age.

Figure 18. Kernel age density estimation for ECLS-B 9-month national data: 2001-02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

Because mental and motor development is explosive during infancy, age and infant 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 are first clustered by age group, and the mean and standard deviation, representing the ability distribution in each age group, are then used to condition group member ability estimates. The gains in precision

obtained with multi-group IRT are thought to be slight, but help ensure consistency when individual observations are scored.

To implement multi-group IRT with ECLS-B data, observations were classified into one of nine age groups between 8 and 16 months of age. There were insufficient observations with ages outside this range to form their own age groups. These older- and younger-age observations were added to the nearest age group at either end of the classification. Observations in the 12-month age group were selected as the standard of reference for item calibration purposes, the same age group used for publisher item calibrations. This age group is assumed to be normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$, $N(0, 1)$. During item calibration, other age-groups and items find their positions along the ability scale in relation to this reference population.

Multi-group IRT has been applied to ECLS-B item calibrations using Bilog-MG and in-house software. This first set of software represents an industry standard and is useful for assessing the precision and accuracy of results. In-house software provides better graphics for visual inspection of item fit, together with almost unlimited flexibility during test equating and analysis. The two sets of software use multi-group IRT and produce results that are essentially identical.

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 because the measurement process is objective, external to either the specific set of items or people found on any given occasion.

In ideal circumstances, ECLS-B and publisher item calibrations yield pairs of IRT item difficulty parameters that are related by means of a simple linear transformation of origin and scale. With this simple transformation, BSF-R ability estimates can be reported using the BSID-II scale metric, providing test results that are comparable from one testing occasion to the next.

When the values of item difficulty parameters obtained from independent calibrations of different samples are plotted on a graph in two dimensions, the resulting data points should align themselves along a straight line. This straight line provides evidence of IRT parameter invariance properties. A straight line shows that the item calibrations define comparable intervals of difficulty along the full extent of either scale.

A perfect straight line would show that any two items differ in difficulty by the same relative amount no matter which sample of subjects is used to calibrate the items. Similarly, any two subjects would differ in relative ability by the same amount no matter which instrument is used. Either scale would produce the same hierarchy of substantive differences between items and between subjects. These relationships enable a simple linear transformation of origin and scale that translates scores on one test into the scale metric of the other test.

Examining results from different samples of items and people obtained on successive occasions shows the degree to which the response data satisfy IRT assumptions. Item parameters are calibrated separately for each of these occasions and are then examined for evidence of parameter invariance properties. If the evidence is found to be supportive, then objective measurement has been obtained. Evidence of objective measurement allows ECLS-B to report results on a common publisher scale metric spanning a broad range of infant development. ECLS-B data provide an opportunity to examine the BSID-II and BSF-R for evidence of objective measurement.

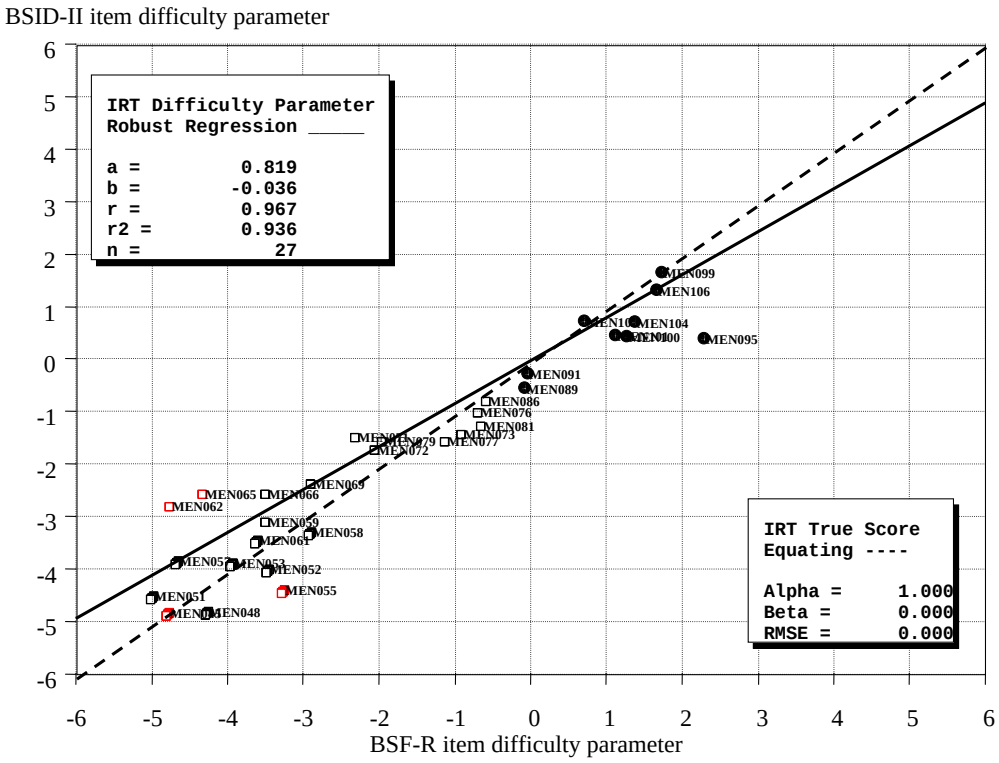
4.5 Nine-Month BSF-R Mental Scale

IRT item difficulty parameters for the mental scale, based on separate calibrations of publisher and ECLS-B item responses, are shown in figure 19. Values of BSID-II item difficulty parameters *y* are aligned along the figure's vertical axis, while BSF-R item difficulty parameters *x* are aligned along the horizontal axis. In other words, the diagonal that goes from the lower left hand corner across to the upper right hand corner represents publisher data. The line that is tilted slightly off the diagonal is ECLS-B data. This figure demonstrates how ECLS-B GSF-R items match up against publisher items in terms of difficulty parameter. The best linear estimate of the relationship between the two sets of item difficulty parameters is reported in the box at the upper left of figure 19:

$$y = ax + b = 0.819x - 0.036,$$

where publisher item difficulty parameter *y* is expressed as a function of the ECLS-B item difficulty parameter *x*. This is a simple regression, where publisher difficulty parameter values have been regressed on the corresponding ECLS-B values. The resulting regression line is represented by a black line in figure 19.

Figure 19. BSF-R mental scale: IRT item difficulty parameter *b* calibrated separately using BSID-II standardization sample and ECLS-B 9-month data: 1993 and 2001-02



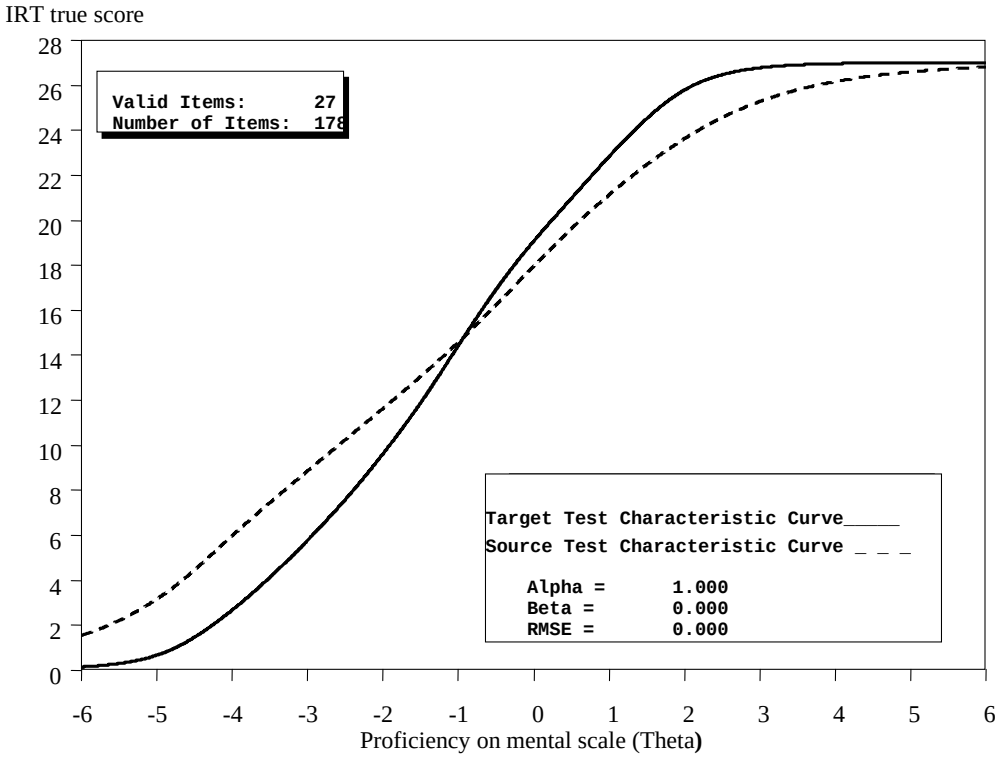
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Visual inspection shows that there is appreciable dispersion in IRT difficulty parameter values around the line of central tendency. An r^2 of .94 implies that about 94 percent of the total variance is common to both scales. Item difficulty parameters that are especially far removed from the regression line show evidence that parameter invariance properties do not apply. Figure 19 shows considerable dispersion in item difficulty parameter values at the low end of the scale, where the basal items are concentrated.

An examination of item difficulty parameters *b* provides a useful heuristic, showing how the original BSID-II and BSF-R scale metrics are related. However, the relationship shown in figure 20 ignores item discrimination parameters *a*. IRT true score equating (Lord and Stocking 1983) simultaneously considers both item difficulty parameters *b* and item discrimination parameters *a*. True

score equating is based on the test characteristic curve (TCC). The TCC is the sum of the ordinates of the item characteristic curves (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. In fact, the publisher used this same metric when scoring the standardization data for the BSID-II.

Figure 20. BSF-R mental scale: Test characteristic curves for publisher and ECLS-B data before true score equating: 1993 and 2001-02

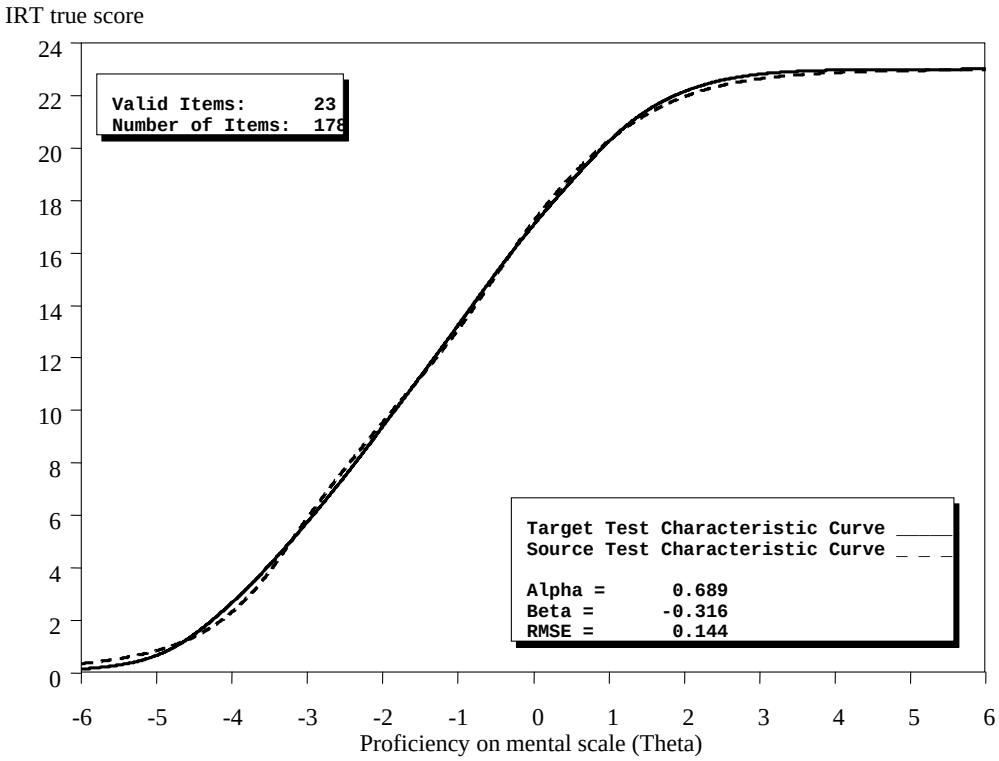


SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

IRT item parameters are used to generate values for publisher and ECLS-B TCCs across all levels of ability, as seen in figure 21. The linear transformation that minimizes the vertical distance between the two curves across all ability levels is used to place ECLS-B parameter estimates on the same metric established with publisher item calibrations.

Equating coefficients α and β (alpha and beta) provide the linear transformation of origin β and scale α that best aligns the ECLS-B (source) TCC with the publisher (target) TCC, based on items that are common to both scales. The alignment of true score values, after true score equating, is shown in figure 21. Comparing figure 21 with 22, the equated test scores appear fairly consistent, and acceptable consistency is found across the entire range of the ability distribution as represented by the nearly overlapping of the curves.

Figure 21. BSF-R mental scale: Test characteristic curves for publisher and ECLS-B data after true score equating: 1993 and 2001-02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; publisher standardization data set of the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

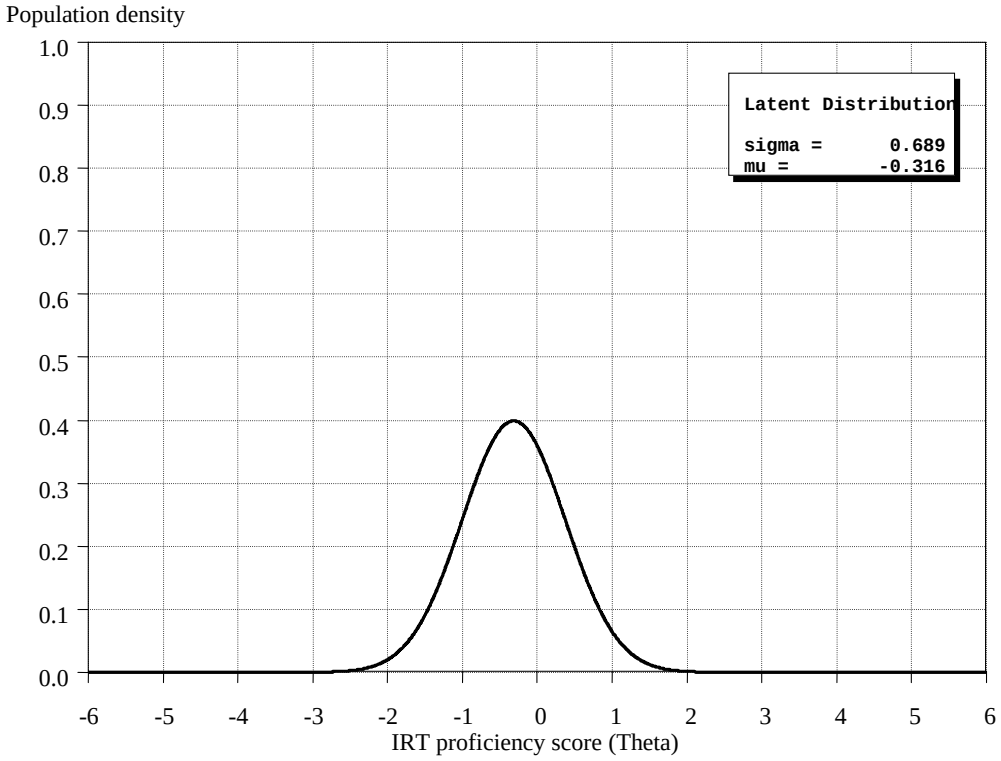
The new equating coefficients are also reported in the box found at the lower right corner of figure 21. The relationship:

$$y = \alpha x + \beta = 0.689x - 0.316,$$

is found to provide a better representation of the relationship between the two scale metrics. In figure 21, the solid black line represents the linear relationship obtained after true score equating using the TCC as

the standard of comparison. IRT true score equating uses both the 2PL *a* and *b* item parameters and provides the best estimate of the relationship between publisher and BSF-R scales. In this particular case, the results obtained with true score equating are quite similar to the initial values obtained by regression, using only the IRT difficulty parameters.

Figure 22. 9-month BSF-R mental scale: ECLS-B 12-month ability distribution in publisher metric after true score equating: 1993 and 2001-02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

True score equating repositions the ECLS-B latent ability distribution on the IRT scale established with publisher data. ECLS-B ability scores are placed on publisher scale metric by applying the same α and β equating coefficients to ECLS-B ability estimates θ (theta). Since ECLS-B calibrations have been arbitrarily set to produce a $N(0, 1)$ distribution, with a mean of $\mu = 0$ and standard deviation of $\sigma = 1$ at 12-months of age, after true score equating, the same distribution will have mean $\mu = \beta = -0.316$ and standard deviation $\sigma = \alpha = 0.689$, as shown in figure 22. The reference population for the new scale metric is based on 12-month-old infants found in publisher data, currently the best available standard for this age group. Thus, a mean of $\overline{X} = \alpha = -0.316$ implies that the 12-month-old national sample is

centered just below the median obtained for 12-month-old infants with publisher data. Additionally, the mental development of 12-month-old infants found in the ECLS-B sample is appreciably less varied than that of 12-month-olds found in the publisher's standardization sample.

The age-ability relationship for both ECLS-B and publisher standardization is illustrated with group means shown in figure 23. The group means are close to the publisher standard, but show that with increasing age the group means fall progressively short of the publisher standard. It is as if these scores are lined up on an inclined plane that does not tilt upward as high as that of the publisher's standardization data set. This pattern does not indicate that the progressively older children in the ECLS-B are less able than younger ECLS-B children. On average, they are much more able than younger children. What the standardized ECLS-B scores show is that at progressively older ages, the children in the ECLS-B sample are less developed than other children of the same age (i.e., the children in the publisher's standardization sample)[1]. It is not possible to define a particular age beyond which the scores decline, as if there were a threshold. Rather the scores show a general, steady decline from 8 months and beyond.

Since these data are cross-sectional, this does not represent a longitudinal trend involving individual children over time. Instead, it suggests that older-age children in the ECLS-B sample are increasingly less developed relative to other children of the same age found in the publisher's standardization data set. The difference is real and reflects levels of performance on individual tasks. Analysts should be mindful that there are many individual observations in the ECLS-B sample outside the range of these group means. At present, the study cannot explain why ECLS-B children who were interviewed late are less developed mentally than children of the same age.

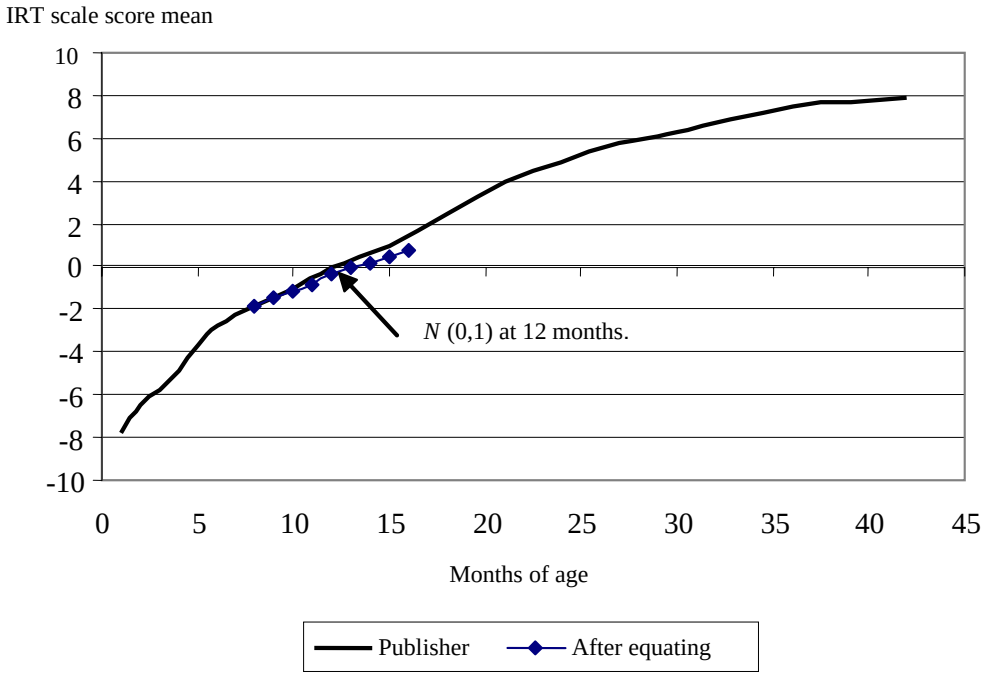
For the ECLS-B, true score equating was carried out in conjunction with an analysis of differential item functioning (DIF). The objective of the analysis was to exclude from equating any items that functioned differently in publisher and ECLS-B calibrations. The removal of items with significant DIF is designed to yield comparable measures on either test.

For this purpose, a parametric IRT measure of differential test functioning (DTF) for the entire test was used (Raju, van der Linden, and Fleer 1995). With this procedure, DIF values for

[1] These comparisons are with the publisher standardization data set, which is the standard against which BSF-R scores should be evaluated. The goal is to determine whether the BSF-R "is the same as" the BSID-II. Subsequent exploratory analyses to understand these observed differences, as well as longitudinal analyses and recalibration of the 9-month and 2-year BSF-R scores has shown that these observed differences are even greater at 2-years and are not due to bias or scaling problems, and are not artifacts.

individual items sum to the total DTF. This makes it possible to determine how test bias will change after removing an item from the test. Unlike other measures, this DIF index does not assume that the other items in the test are unbiased. The DIF analysis assumes that the two sets of item parameters have been calibrated on a common scale. This suggests a procedure for concurrent equating and DIF analysis that will select a subset of items that matches the target test as closely as possible.

Figure 23. 9-month mental scale: Mean age-ability relationship in ECLS-B and publisher data: 1993 and 2001-02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set of the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Initially, publisher and ECLS-B items are calibrated independently with different scale metrics. True score equating aligns the BSF-R (source) test with the publisher (target) test. DIF analysis follows, using the BSF-R items representing the focal group, and publisher items representing the reference group. DIF values are calculated for each item, and the values are summed to represent total test DTF. If the item with the largest DIF index exceeds a certain value, this item is temporarily removed from the scale only for the purpose of estimating this transformation. Items continue to be excluded until there are no remaining items exceeding this value. DIF analysis, together with the removal of differentially functioning items, continues until an acceptably small value of total test DTF is obtained.

Once this has been achieved, there can be confidence that a more refined or stable set of transformation constants has been obtained.

This procedure is useful because it not only identifies items that show evidence of bias, but also provides an estimate of the effects of item removal on total test DTF. This does not mean that the final selection of items is without bias. Rather, the biases of the remaining items compensate one another in such a way that the overall test is relatively unbiased across the entire range of the ability distribution.

In practice, the procedure will often identify items that show dependencies due to similar phrasing or content. In the BSF-R, these dependencies were found among several of the cube or block items in the mental scale, where infants were asked to place a certain number of cubes in a cup or to retain the cubes in a cup for a few seconds. In the motor scale, DIF was similarly detected among some of the standing and walking items. Not only did some of these items show similarity in content, but often one item had to be successfully completed before a related item could even be attempted. This introduces contingencies that violate the IRT principle of independence among items.

True score equating, implemented in conjunction with DIF analysis, excluded 4 items from the initial set of 31 BSF-R mental items during equating but were retained for scoring. True-score equating simply places BSF-R results on the same scale metric used by the publisher to determine standardized scores. In the case of the BSF-R, true-score equating revealed inconsistencies between the BSF-R and publisher scales involving a few of the items. The purpose of equating is to put BSF-R scores on the publisher metric. In these circumstances, it is usually best to exclude such items while establishing the appropriate transformation in scale metric. The improved fit is likely to result in a better estimate of the origin and scale used in the transformation. However, within the ECLS-B data, these items were included in the scoring because they strengthen the overall scale scores and fit the ECLS-B data well.

On the other hand, items that show evidence of DIF should probably be retained in the scale for purposes of scoring. Item calibrations reveal that BSF-R items fit ECLS-B data very nicely and thus should be considered as part of the scale. When scoring, these items increase the precision of ability estimates and ultimately they enhance scale reliabilities. Issues of scale content and construct validity also provided additional justification for retaining the items in the scale.

The item composition of the final mental scale is reported in table 7, where the items that were either excluded from the scale or excluded during equating are noted. The items are listed in the

order of difficulty recommended by the publisher. IRT difficulty parameters *b* and discrimination parameters *a* have been generated for all items, however, DIF analysis shows that it would be inappropriate to base the equating on some of these items. Items that have been excluded from equating are as noted in the rightmost column. Item difficulty parameters are found in the range of $-4 < b < 1.2$. There are several basal and ceiling items with $a > 1$, but few core items show this much discrimination.

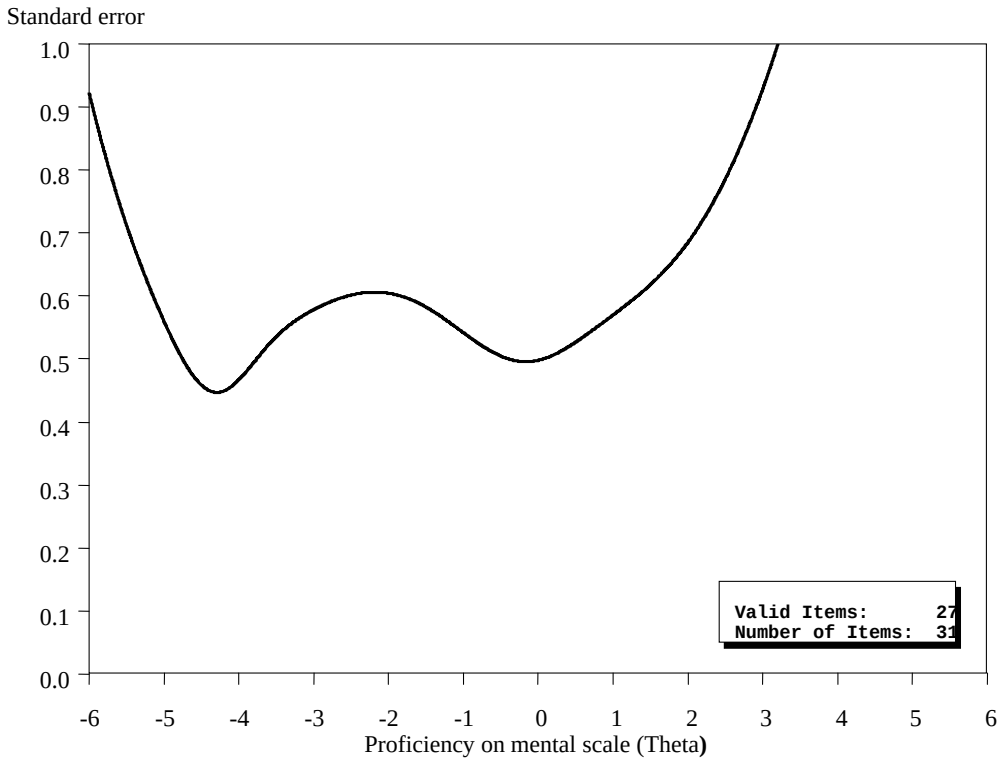
Table 7. 9-month IRT difficulty parameter *b* and discrimination parameter *a* for items comprising the BSF-R mental scale: IRT 2-PL item calibrations using ECLS-B national data, after true score equating: 2001–02

Item	Item label	Item set	IRT difficulty	IRT discrimination
MEN045	Picks up cube[1]	Basal	-3.67	0.85
MEN048	Plays with string	Basal	-3.31	2.66
MEN051	Regards pellet	Basal	-3.81	1.67
MEN052	Bangs in play	Basal	-2.75	1.54
MEN053	Reaches for second cube	Basal	-3.09	0.62
MEN055	Lifts inverted cup[1]	Basal	-2.61	0.93
MEN057	Picks up cube deftly	Basal	-3.58	0.96
MEN058	Retains two cubes for 3 seconds	Basal	-2.36	1.17
MEN059	Manipulates bell, showing interest in detail	Core	-2.77	0.36
MEN061	Vocalizes three different vowel sounds	Basal	-2.86	0.60
MEN062	Pulls string adaptively to secure ring[1]	Core	-3.65	0.40
MEN065	Retains two of three cubes for 3 seconds[1]	Core	-3.35	0.24
MEN066	Rings bell purposely	Core	-2.77	0.60
MEN069	Looks at pictures in book	Core	-2.36	0.51
MEN071	Repeats vowel-consonant combinations[2]	Core	-1.95	0.63
MEN072	Looks for contents of box	Core	-1.77	1.03
MEN073	Turns pages of book[2]	Core	-0.98	0.68
MEN076	Jabbers expressively	Core	-0.84	0.70
MEN077	Pushes car	Core	-1.13	0.90
MEN079	Fingers holes in pegboard	Core	-1.71	0.43
MEN081	Responds to spoken request	Core	-0.80	0.47
MEN086	Puts three cubes in cup	Core	-0.76	1.27
MEN089	Puts six beads in box	Ceiling	-0.41	1.72
MEN091	Scribbles spontaneously	Ceiling	-0.38	1.25
MEN095	Puts nine cubes in cup	Ceiling	1.23	0.79
MEN099	Points to two pictures	Ceiling	0.85	0.80
MEN100	Uses two different words appropriately	Ceiling	0.53	1.08
MEN101	Shows shoes, other clothing, or object	Ceiling	0.42	0.97
MEN102	Retrieves toy (visible displacements)[2]	Ceiling	0.14	0.53
MEN104	Uses rod to attain toy	Ceiling	0.60	0.60
MEN106	Uses word(s) to make wants known[2]	Ceiling	0.80	1.16

[1] Item excluded from the ECLS-B BSF-R scale.
[2] Item excluded from equating with publisher scale, but used in scoring the ECLS-B BSF-R scale.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation., 1993.

Items not well represented by the scale include difficult items answered correctly by many individuals of low ability. Individuals not well represented by the scale include able infants who miss several easy items. This provides a basis for assessing the overall quality of the scale, including measures of scale reliability and a standard error of measurement for each examinee. Figure 24 summarizes calibration results obtained with the remaining 27 mental BSF-R items.

Figure 24. BSF-R mental scale: Standard errors for ECLS-B 9-month ability estimates after true score equating: 2001-02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

The standard error of measurement is based on IRT item parameters calibrated using ECLS-B item responses, after the BSF-R items have been equated to publisher scale metric. For the ECLS-B, the objective was to obtain standard errors below 0.4 standard deviations for the 9-month population well out into the tails of the ability distribution. Simulations showed that this level of error would correspond with an IRT reliability coefficient of r_{xx} = .80, which consultants to ECLS-B considered satisfactory. The figure shows that this performance objective was attained with the basal and ceiling item sets. This objective was nearly achieved by the core item set by 0.2 standard deviations (see figure 24 in the range of -3 to -1 theta).

The IRT reliability coefficient obtained for the BSF-R mental scale in ECLS-B is $r_{xx} = 0.80$.[2] The corresponding reliability coefficient for the same subset of items using publisher calibrations is $r_{xx} = .87$. These coefficients represent the reliability of the BSF-R scale when used with only the 9-month target population. Another way to assess the reliability of the BSF-R mental scale is to examine the average correlation between all possible pairs of plausible values for all observations in the ECLS-B sample. The correlation among plausible values is analogous to the correlation among parallel forms, which serves as the foundation for the assessment of reliability in classical test theory. Five plausible values are drawn within the error distribution for each observation to provide a random Monte Carlo representation of the error component in the BSF-R mental infant development scale. The average correlation among all possible pairs of plausible values for this scale is $r_{xx} = .79$.

4.6 Nine-Month BSF-R Motor Scale

IRT item difficulty parameters for the motor scale, based on separate calibrations of publisher and ECLS-B item responses, are shown in figure 25. The best linear estimate of the relationship between the two sets of item difficulty parameters is reported in the box at the upper left of the figure:

$$y = ax + b = 0.886x - 0.187$$

where publisher item difficulty parameter y is expressed as a function of the ECLS-B item difficulty parameter x. The resulting regression line is represented by a black line in figure 25.

Once again, there is appreciable dispersion in IRT difficulty parameters around the line of central tendency. The coefficient $r^2 = .92$ shows that about 92 percent of the total variance is common to both scales. This is about the same as the mental scale and often encountered in an exercise of this type.

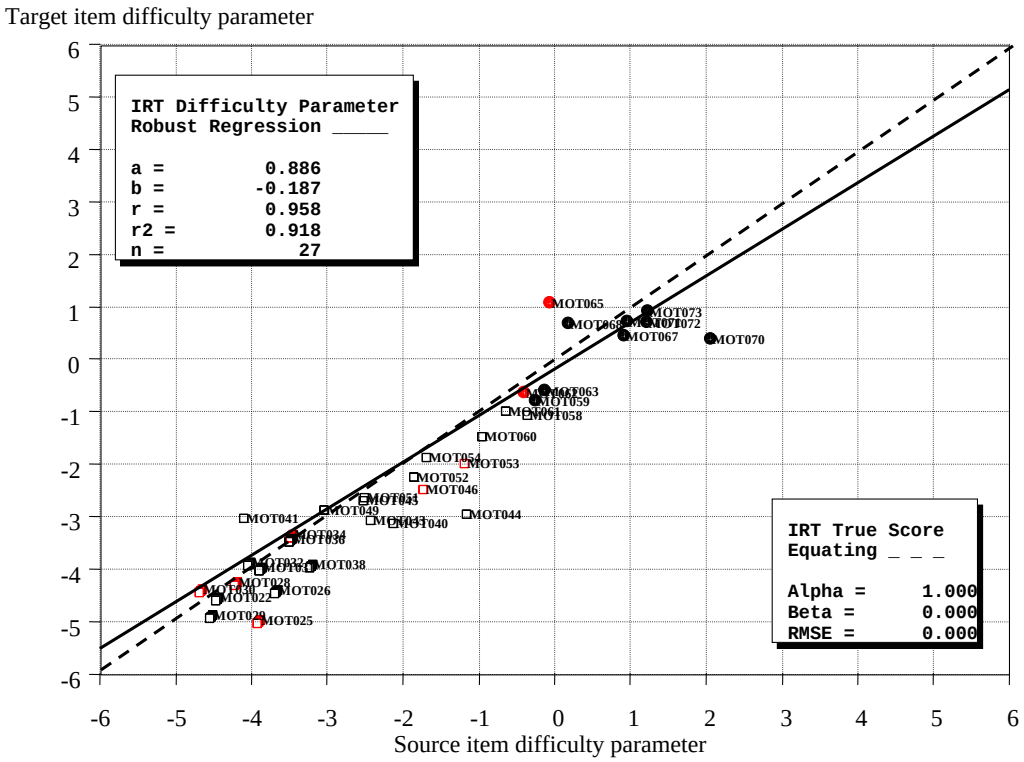
---

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

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

where the $A(X_k)$ are normal ordinate weights for quadrature points X_k spanning the distribution of ability for 9-month infants, with $\sum_{k=1}^{q} X(\theta_k) = 1$.

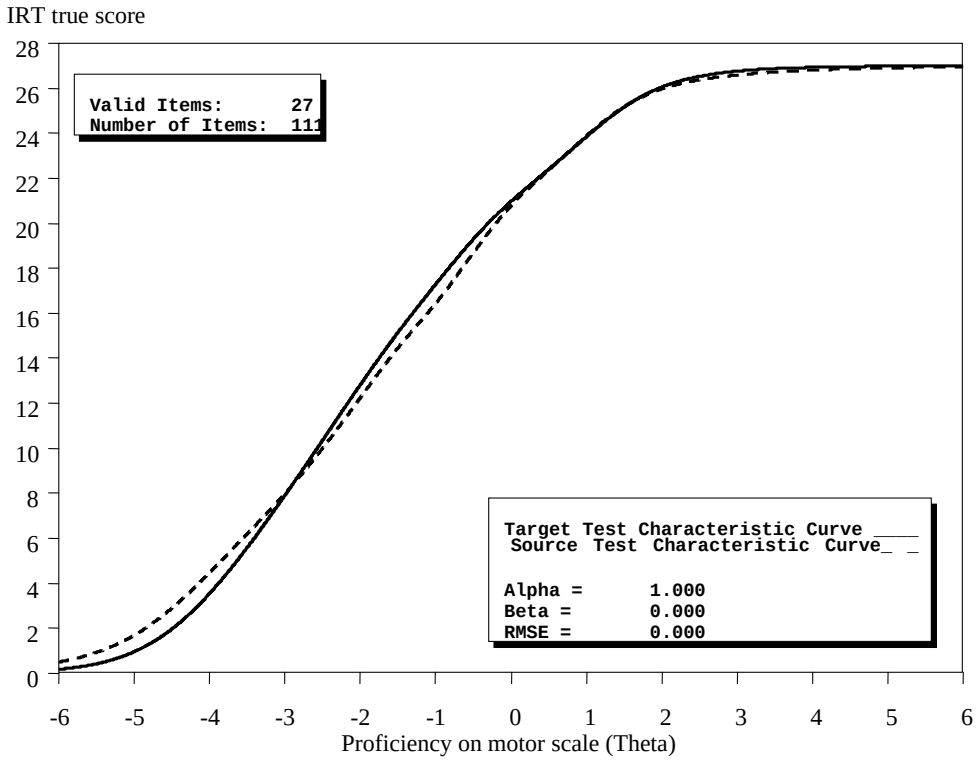
Figure 25. BSF-R motor scale: IRT item difficulty parameter *b* calibrated separately using publisher and ECLS-B data" 1993 and 2001-02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; publisher standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

IRT true score equating, conducted in conjunction with item DIF analysis, provides a better method for placing BSF-R results on publisher metric. The test characteristics curves for BSF-R items on both scales, prior to equating, are shown in figure 26, where reference lines show the center of the ECLS-B population at $\theta = 0$ and the corresponding point on the publisher scale. The linear transformation that minimizes the vertical difference between the two test characteristic curves across all ability levels is used to place ECLS-B parameter estimates on the same metric established with publisher item calibrations.

Figure 26. 9-month BSF-R motor scale test characteristic curves (TCCs) for publisher and ECLS-B data before true score equating: 1993 and 2001-02



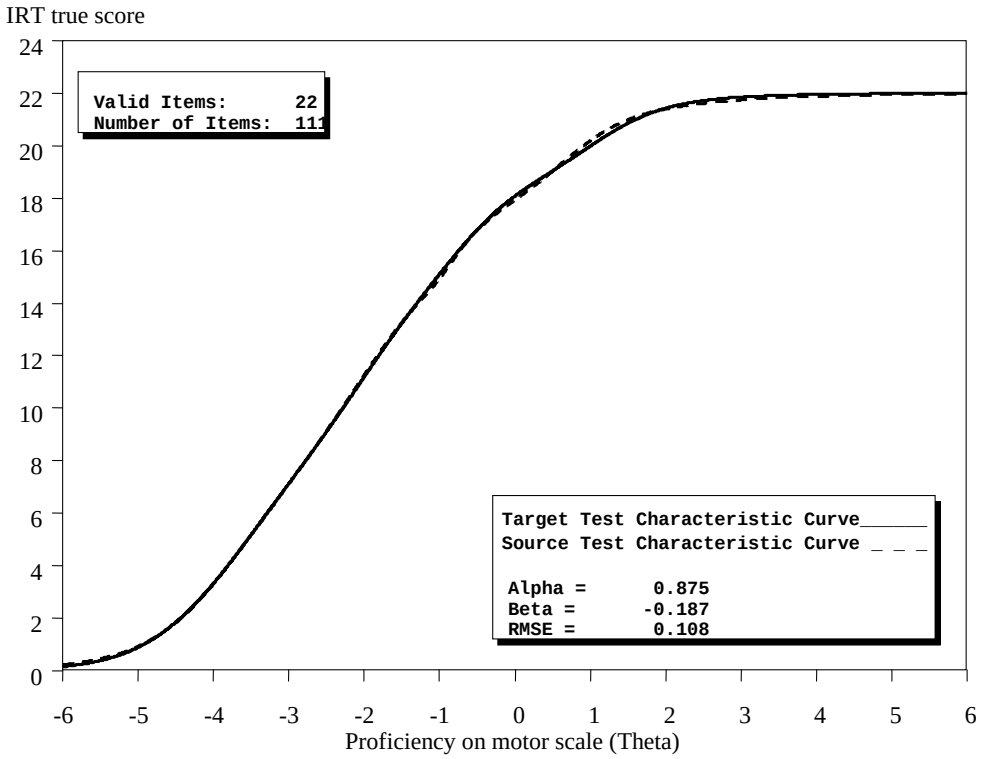
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; publisher standardization data set of the Bayley Scales of Infant Development, Second Edition (BISD-II) The Psychological Corporation, 1993.

The alignment of true score values, after true score equating, is shown in figure 27. The equated test scores are fairly consistent, and respectable consistency is found across the entire range of the ability distribution. The relationship:

$$y = \alpha x + \beta = 0.875x - 0.187$$

is found to provide a better representation of the relationship between the two scale metrics.

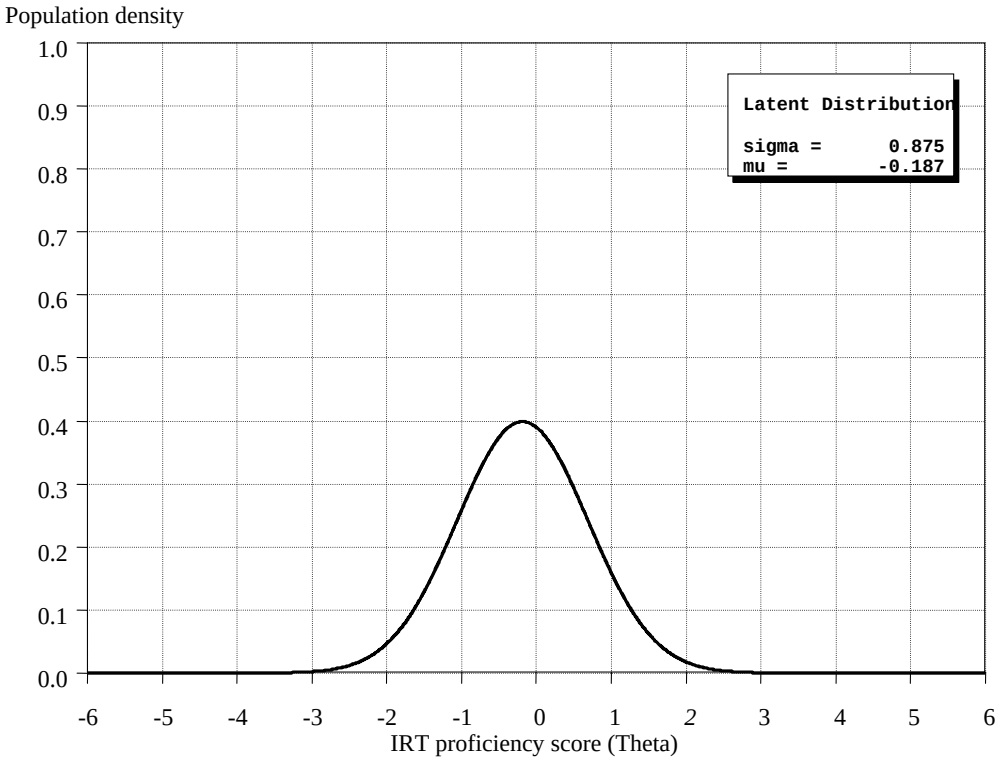
Figure 27. 9-month BSF-R motor scale test characteristic curves (TCCs) for publisher and ECLS-B data after true score equating: 1993 and 2001–02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

ECLS-B ability scores are placed on publisher scale metric by applying the same α and β equating coefficients to ECLS-B ability estimates θ (theta). Since ECLS-B calibrations have been arbitrarily set to produce a $N(0, 1)$ distribution, with a mean of $\mu = 0$ and standard deviation of $\sigma = 1$ at 12-months of age, after true score equating, the same distribution will have standard deviation $\sigma = \alpha = 0.875$ and mean $\mu = \beta = -0.187$, as shown in figure 28.

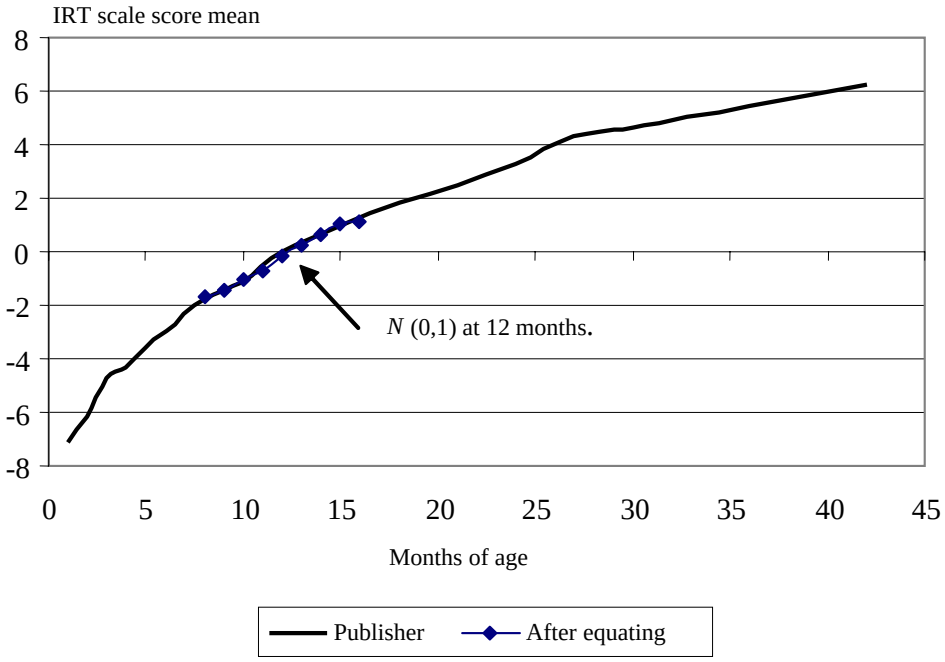
Figure 28. 9-month BSF-R motor scale: ECLS-B 9-month ability distribution in publisher metric after true score equating: 2001–02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set for the Bayley Scales of Infant Development, Second Edition (BSID-II) The Psychological Corporation, 1993.

True score equating, in conjunction with DIF analysis, removed 4 of the initial set of 35 BSF-R motor items to establish the transformation of scale used in equating, although these items are retained for scoring purposes. The central tendency of age-ability relationships is shown in figure 29, where ECLS-B mean ability estimates have been superimposed on publisher data. ECLS-B group means are very close to those found in the publisher's standardization data set, and there is no evidence of trend in the ECLS-B sample, involving children of older age.

Figure 29. 9-month BSF-R motor scale: Mean age-ability relationship in ECLS-B and publisher data: 1993 and 2001–02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set of the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Items considered for the BSF-R motor scale are reported in table 8, where they are listed in the order of difficulty recommended by the publisher. Items that were either excluded from the final scale or not used during equating have been noted in the last column of the table. Item difficulty parameters are found in the general range of $-4.1 < b < 1.6$. All three BSF-R item sets contain items with discrimination parameter $a > 1$, implying that the scale yields considerable information across the entire ability distribution. The problem with the motor scale is that there are several items with $a > 2$, suggesting that there may be dependencies between specific pairs of items found at the same or similar difficulty levels. If one could imagine a placid pool of water into which a number of small pebbles and two large stones are thrown simultaneously, the resulting ripples of the larger stones would "wash out" the ripples arising from the pebbles. By eliminating one of the large stones, the ripples from the smaller pebbles may still be seen. Similarly, two highly discriminating items that are close together in terms of their ability level would "wash out" the contribution of the less discriminating items, making it impossible to judge the relative contributions of each. Indeed this was the motive for removing some of the more highly discriminating items from the final scale.

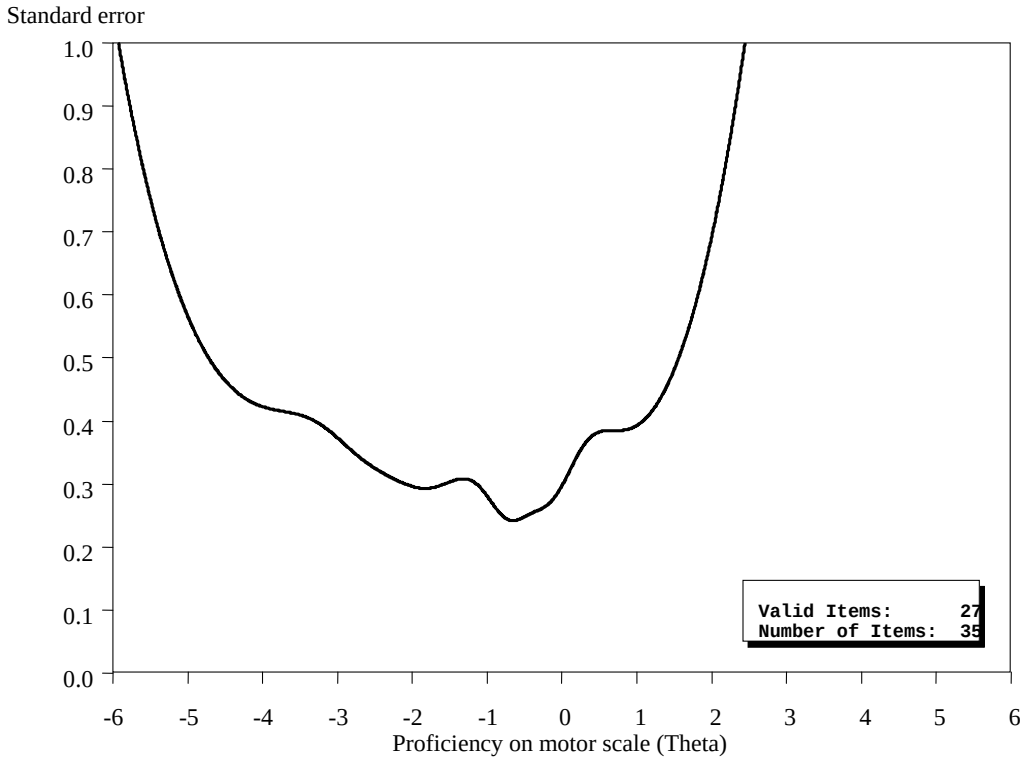
Table 8. 9-month IRT difficulty parameter *b* and discrimination parameter *a* for items comprising the BSF-R motor scale: IRT 2-PL item calibrations using ECLS-B national data, after true score equating: 2001–02

Item	Item label	Item set	IRT difficulty	IRT discrimination
MOT022	Sits with slight support for 10 seconds	Basal	-4.10	1.22
MOT025	Shifts weight on arms[1]	Basal	-3.63	0.59
MOT026	Turns from back to side	Basal	-3.43	0.97
MOT028	Sits alone momentarily[1]	Basal	-3.89	1.21
MOT029	Uses whole hand to grasp rod	Basal	-4.17	1.17
MOT030	Reaches unilaterally[1]	Basal	-4.29	0.96
MOT031	Uses partial thumb opposition to grasp cube	Basal	-3.60	0.99
MOT032	Attempts to secure pellet	Basal	-3.73	1.47
MOT034	Sits alone for 30 seconds[1]	Basal	-3.25	1.32
MOT036	Sits alone steadily	Basal	-3.26	1.21
MOT038	Turns from back to stomach	Basal	-3.03	1.11
MOT040	Makes early stepping movements[2]	Core	-2.08	1.58
MOT041	Uses whole hand to grasp pellet	Core	-3.79	0.62
MOT043	Moves forward using prewalking methods	Core	-2.33	1.61
MOT044	Supports weight momentarily[2]	Core	-1.24	1.89
MOT045	Pulls to standing position	Core	-2.43	0.87
MOT046	Shifts weight while standing[1]	Core	-1.73	1.28
MOT049	Uses partial thumb opposition to grasp pellet	Core	-2.87	0.62
MOT051	Moves from sitting to creeping position	Core	-2.41	2.01
MOT052	Raises self to standing position	Core	-1.84	2.09
MOT053	Attempts to walk[1]	Core	-1.27	1.52
MOT054	Walks sideways while holding on to furniture	Core	-1.70	2.43
MOT058	Grasps pencil at farthest end	Core	-0.54	0.69
MOT059	Stands up I	Ceiling	-0.46	1.90
MOT060	Walks with help	Core	-1.07	2.03
MOT061	Stands alone	Core	-0.79	3.91
MOT062	Walks alone[1]	Ceiling	-0.59	2.92
MOT063	Walks alone with good coordination[2]	Ceiling	-0.36	3.49
MOT065	Squats briefly[1]	Ceiling	-0.30	1.30
MOT067	Walks backward	Ceiling	0.56	1.85
MOT068	Stands up II[2]	Ceiling	-0.09	1.25
MOT070	Grasps pencil at middle	Ceiling	1.55	0.51
MOT071	Walks sideways[2]	Ceiling	0.60	1.70
MOT072	Stands on right foot with help	Ceiling	0.81	1.37
MOT073	Stands on left foot with help	Ceiling	0.83	1.38

[1] Item excluded from the ECLS-B BSF-R scale.
[2] Item excluded from equating with publisher scale, but used in scoring the ECLS-B BSF-R scale.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

Figure 30, showing the standard error of measurement across the ability distribution, summarizes calibration results obtained with the 27 motor items retained in the BSF-R scale. The IRT reliability coefficient obtained for the BSF-R motor scale used in ECLS-B is r_{xx} = .92 for the sample as a whole. The reliability coefficient for the same subset of publisher items is r_{xx} = .87. The average correlation among all possible pairs of plausible values for the BSF-R motor scale is r_{xx} = .92.

Figure 30. BSF-R motor scale: Standard errors for ECLS-B 9-month ability estimates after true score equating: 2001-02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

4.7 Differential Test Functioning and Differential Item Functioning

The 9-month BSF-R assessments were examined for evidence of differential test functioning (DTF) and differential item functioning (DIF) in the ECLS-B. This involves comparisons of test performance between a focal group (e.g., African-American infants) and a reference group (e.g. White infants) once the individuals in the two groups have been matched or "blocked" on their ability estimates. It is not expected that the different subgroups will perform identically on the same test. Rather, infants

from the different groups who are matched in terms of their overall ability should have the same probability of obtaining correct responses to the set of items. There should be no relative advantage or disadvantage in obtaining correct responses based on the infants' subgroup membership.

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. Item response theory (IRT) provides a unified framework for investigating issues of statistical bias at both the item and test 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 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. Each file was then scored separately, using an identical set of BSF-R item parameters. The scoring effectively classifies each observation by ability level. As each observation is scored separately in each group, marginal likelihoods are accumulated for each item score across all levels of ability. When all observations have been scored, a new set of IRT parameters are fitted to the marginal likelihoods in a single iteration. The new sets of item parameters represent the response characteristics for each respective focal or reference group across all levels of ability.

The issue to be addressed in DTF analysis is whether infants *at the same level of ability* will on average obtain the same number-right score on the same test. This issue is addressed in IRT by comparing the test characteristic curves (TCCs) for the two groups. Any misalignment of TCCs provides evidence of DTF. The total number-right score at each level of ability is examined by comparing IRT true-scores for each focal and reference group comparison.

The new set of parameter estimates is used for this purpose. The TCCs of the focal (source) and reference (target) tests are compared across all levels of ability. The weighted sum of squared deviations separating the source and target tests is used as a DTF index. The DTF coefficient quantifies the degree of misalignment between the two curves, expressed in squared raw score units. The square root of the DTF coefficient is a root mean squared error (RMSE), expressed in raw score units. In interpreting the magnitude of RMSE values, one should bear in mind the maximum raw score possible or an average raw score on the test in question. These measures of dispersion around the target test characteristic curve are the DTF statistics reported in the literature.

However, from a practical point of view, the analysis is generally more concerned with the overall magnitude and direction of statistical bias as this affects the reported ability estimates. With large samples such as the ECLS-B, virtually any DTF coefficient will be statistically significant. This merely implies that it is appropriate to generalize from the focal and reference group samples to the ECLS-B population and affirm that at least some differential test function exists when these instruments are used with different subgroups of the population. More importantly, assuming that at least some statistical bias exists in the population, then what is the overall direction of statistical bias and is this bias important in magnitude?

Thus, it is 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 are obtained with IRT true-score equating, which shows the linear transformation of origin and scale required to align the source (focal) and target (reference) tests. In the context of DTF analysis, equating constants α (slope) and β (origin) are expressions of the overall statistical bias to be expected when the assessment instrument is used with the focal group.

Both equating coefficients are reported in population standard deviation units and are 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 is that slope $\alpha = 1$ and origin $\beta = 0$, indicating that no transformation is warranted when comparing the two groups. Although these coefficients are not likely to be reported in the DTF literature, conceptually they are very useful and easy to understand.

DTF statistics for the ECLS-B at 9 months are reported in table 9 for nine focal and reference group comparisons. With the large sample size available in ECLS-B, virtually all DTF and RMSE measures prior to equating are statistically significant, whereas virtually no measure of dispersion between TCCs is statistically significant once α and β are used to relate the two groups. This demonstrates rather conclusively that the main difference between focal and reference groups on BSF-R instruments in the ECLS-B is uniquely identified as a distinctly linear form of statistical bias. At some risk of simplification, the β coefficient aptly summarized the magnitude of this statistical bias, conveniently expressed in population standard deviation units.

Table 9. 2-year BSF-R differential test function analysis coefficients

Focal Group – Reference Group	Coefficient	Mental	Motor
African American – White			
	DTF	0.123	0.020
	RMSE	0.351	0.141
	Alpha	1.067	1.027
	Beta	-0.012	0.018
Hispanic – White			
	DTF	0.077	0.008
	RMSE	0.277	0.091
	Alpha	1.033	1.017
	Beta	-0.051	0.013
Asian – White			
	DTF	0.674	0.018
	RMSE	0.821	0.133
	Alpha	1.183	1.024
	Beta	0.045	0.027
Female – Male			
	DTF	0.080	0.008
	RMSE	0.283	0.087
	Alpha	0.995	0.999
	Beta	0.093	0.017
Low – High socioeconomic status			
	DTF	0.029	0.010
	RMSE	0.171	0.101
	Alpha	0.990	1.012
	Beta	-0.067	0.007
Premature – Full term			
	DTF	1.777	0.322
	RMSE	1.333	0.567
	Alpha	1.190	1.045
	Beta	-0.119	-0.049
A (avoidant) – B (secure) Toddler Security of Attachment			
	DTF	0.015	0.008
	RMSE	0.122	0.087
	Alpha	1.028	1.012
	Beta	0.009	0.015

See note at end of table.

Table 9. 2-year BSF-R differential test function analysis coefficients—Continued

Focal Group – Reference Group	Coefficient	Mental	Motor
C (anxious/resistant) -- B (secure) Toddler Security of Attachment			
	DTF	0.271	0.017
	RMSE	0.520	0.131
	Alpha	1.085	1.013
	Beta	-0.041	0.009
D (disorganized) -- B (secure) Toddler Security of Attachment			
	DTF	0.145	0.005
	RMSE	0.381	0.071
	Alpha	1.068	1.001
	Beta	-0.028	-0.011

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine Month Data Collection, 2001–02.

Differential item functioning (DIF) of BSF-R instruments was also examined in ECLS-B. DIF identifies individual items that show an unexpectedly large difference in the probability of a correct response when comparing a focal group (e.g., Black infants) and a reference group (e.g., White infants) once both groups are "blocked" or matched on their total score. An item shows DIF when "individuals having the same ability, but from different groups, do not have the same probability of getting the item right." (Hambleton, Swaminathan, and Rogers 1991).

For the 9-month BSF-R assessments, DIF indices are calculated using parametric IRT procedures developed by Raju, van der Linden, and Fleer (1995). Focal groups examined include African-American, Hispanic, Asian race/ethnicity, female gender, low socioeconomic status (SES), and A-, C- and D-style attachment behavior. The respective reference groups for these comparisons included White race/ethnicity, male gender, high SES, and B-style attachment behavior. Table 10 summarizes these results, showing all of the items on the 9-month BSF-R assessments exhibiting weighted root mean squared errors of 0.10 or greater. Items that showed lower levels of DIF have been excluded from the table.

Table 10. 9-Month BSF-R mental scale and motor scale differential item function

BSID-II Item	Item label	Asian ethnicity	Premature	Type C Anxious-Resistant Attachment
Mental Scale				
MEN059	Manipulates bell, showing interest in detail		0.143	
MEN071	Repeats vowel-consonant combination		0.107	
MEN072	Looks for contents of box		0.121	
MEN099	Points to two pictures			0.119
MEN102	Retrieves toy (visible displacement)	0.159	0.166	
Motor Scale				
MOT036	Sits alone steadily		0.153	
MOT049	Uses partial thumb opposition to grasp pellet		0.128	

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine Month Data Collection, 2001–02.

In sum, four mental items and two motor items show appreciable DIF for children who were born prematurely, but this might be expected, given their at-risk status. One other mental item was found to show appreciable DIF for Asian infants and another showed appreciable DIF for infants who would later be classified as anxious/resistant Type C attachment style behavior.

Inspecting the values of the DTF coefficients in table 9 reveals statistical biases that range from small to minute for the nine focal and reference group comparisons considered in this exercise. These comparisons include three race-ethnicity focal groups (African Americans, Hispanics, and Asians in the ECLS-B population), one comparison each for gender, low-socioeconomic status, premature birth, and three comparisons for toddler security of attachment to be reported later at 24 months of age. Small biases favor female, higher income and to term infants on the BSF-R mental scale. The BSF-R motor scale shows virtually no evidence of statistical bias, except possibly one that favors full term infants. In general, there is little, if any evidence of statistical bias of appreciable magnitude for any of the nine focal groups considered in this DTF analysis.

DTF and RMSE coefficients reported in table 9 are primarily reflections of this same, relatively small statistical bias, now expressed in number-right raw score units. The largest RMSE value is 1.333 score points for premature infants compared with full term infants, but there is also a fairly expressive value RMSE = 0.821 for the Asian—White comparison. These coefficients reveal nothing

about the direction of statistical bias, whereas coefficient β shows that this is negative for premature infants but positive for Asians in the ECLS-B at nine months. Again, as a general rule, all DTF and RMSE coefficients are statistically significant prior to a linear transformation of origin and scale; none are significant following a linear transformation. This shows that the principal difference between focal and reference groups is a question of linear statistical bias. However, it is important to remember that DTF is observed and quantified in lieu of any kind of linear transformation.

4.8 Chapter Summary

The important point of this chapter is that multigroup (i.e., 9 age groups from 8 – 16 months of age) IRT analysis, using the 2-parameter logistic model, were conducted on the mental scale and motor scale of the ECLS-B 9-month national data collection to assess the psychometric qualities of the BSF-R. Item calibrations of ECLS-B and publisher items successfully placed the BSF-R ability estimates on the publisher metric, meaning that the BSF-R scores can be viewed as comparable to the publisher BSID-II scores, as if the full BSID-II had been administered.

Furthermore, IRT analysis determined that both the BSF-R mental and motor scales have good reliability. The goal was to obtain standard errors below 0.4 standard deviations for the 9-month population well out into the tails of the ability distribution. For the BSF-R mental scale as a whole, IRT analysis found an IRT reliability coefficient of r_{xx} = 0.80, which consultants to the ECLS-B considered satisfactory. For the BSF-R motor scale as a whole IRT analysis found an IRT reliability coefficient of r_{xx} = 0.92 for the ECLS-B, which exceeds the reliability coefficient of r_{xx} = .87 for the same subset of BSID-II items in the publisher standardization data set. Lastly, differential item functioning and differential test functioning showed that there was no harmful bias at either the item-level or at the test level.

5. BAYLEY SHORT FORM–RESEARCH EDITION SCORES ON THE PUBLIC-USE FILE

The following section presents the IRT analyses that were conducted in order to develop the BSF-R scores that are in the public-use data file. The rules adopted to ensure reliable scores (i.e., the required minimum number of scored items and the use of basal and ceiling item sets) are first described. This is followed by descriptions of the scores and their derivation. This section then concludes with a table of group differences for each of these scores.

In order to ensure the quality of the BSF-R scores and to conduct these analyses, it was first necessary to determine when a case had sufficient BSF-R items scored in order to generate these scores. Therefore, a decision rule was established in which two-thirds (66 percent) of items in the mental and in the motor scale must be scored as either "C" (credit) or "NC" (no credit) in order to assign a score for each scale. Alternatively stated, no more than one-third of items can be "missing," or not ascertained. Although standard errors provide an appropriate estimate of the precision of Bayley Short Form–Research Edition (BSF-R) ability estimates, National Center for Education Statistics (NCES) staff felt that a certain number of items would have to be answered before a child's ability estimate could be recorded. A decision rule requiring that two-thirds of the items in a scale be answered before a scale score is generated has sometimes been used to ensure adequate content coverage and to address the underlying issue of construct validity. After some deliberation, NCES staff asked that this same rule be applied when scoring the BSF-R.

When using the Bayley Scales of Infant Development, Second Edition (BSID-II) in a clinical setting, the publisher recommends that 90 percent of the items have scores or alternatively that only 10 percent of the items are allowed to be missing. Since scoring the full BSID-II is based on the number-right raw score total, some sort of stringent requirement is needed in order to obtain an appropriate estimate of the child's ability.

These circumstances do not apply in the case of the BSF-R. It is not just that ECLS-B is conducted in a survey context rather than a clinical setting, but also and especially that the number-right raw score plays no role in maximum likelihood scoring with item response theory (IRT). IRT uses only the responses to individual items to estimate the likelihood distribution for each observation. In so doing, maximum likelihood is extremely tolerant of missing data, which affects the standard error but not the central tendency of the ability distribution.

The BSF-R also differs from the BSID-II in another important way. The BSID-II is organized around age-sets of items because it is unreasonable to administer items appropriate for a 36-month-old to an 8-month old. Therefore, for a given age range, say 8 months, a set of items is designated as being appropriate for administration and this appropriate age-set is usually sufficient to obtain a reliable score. However, in those cases where the child scores poorly, the assessor is instructed to administer successively younger age-sets of items until a satisfactory basal measure of performance is achieved. Similarly, in those cases where the child scores quite well, the assessor is instructed to successively administer older age-sets of items until a satisfactory measure of the upper level of performance is achieved. Although the BSID-II manual does not present any statistics about how frequently additional age-sets are necessary, in the large majority of clinical administrations, a single age set of items is sufficient to produce a reliable score.

This approach was not feasible in the ECLS-B, however. Instead, as described, IRT analyses were used to design a core set of items that would be administered to all children. For children who scored between one standard deviation below the mean and one standard deviation above the mean, this core set of items would be sufficient to obtain a reliable score. A (single) basal set of items was designed to be administered to children who scored one or more standard deviation(s) lower than the mean on the core items. Likewise, a (single) ceiling set of items was designed to be administered to children who scored one or more standard deviation(s) or higher than the mean on these core items. Therefore, on the basis of the normal distribution, it was expected that 70 percent of children in the ECLS-B would need only the core sets of items, 15 percent would need the basal set of items and another 15 percent would need the ceiling set of items. On the mental scale, 59.4 percent of the total sample needed only the mental core item set administered, with 5 percent needing the basal item set and 36 percent the ceiling item set. On the motor scale, 63 percent received the motor core items only, with 8 percent needing the basal item set and 29 percent needing the ceiling item set. However, this is for the total sample for the 9-month data collection and does not consider the child's age. Table 11 below) presents the breakdown of core, basal and ceiling item sets by age in months. Only those cases in which the basal or ceiling set needed to be administered are included, i.e., these percentages do not include any cases in which the basal or ceiling set was incorrectly administered.

Finally, the BSID-II 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 is a Mental Development Index score for the mental

scale, and a Psychomotor Development Index score for the motor scale. Similarly, there is no single BSF-R total score. This conforms to standard scoring procedures for the BSID-II.

Table 11. Percent of children by age correctly administered mental and motor scale basal or ceiling item sets, 9-month data collection: 2001–02

		Mental scale		Motor scale	
Age (in months)	N	Basal set (percent)	Ceiling set (percent)	Basal set (percent)	Ceiling set (percent)
Total sample	10,214	5	36	8	29
8	1,616	8.1	12.2	13.6	6.4
9	3,502	5.5	33.0	8.8	15.0
10	2,235	4.1	44.0	5.7	28.5
11	1,166	4.2	45.2	6.0	39.7
12	706	2.5	61.2	4.1	61.9
13	391	1.0	76.2	1.5	78.0
14	240	0.4	78.3	0.8	83.8
15	149	0.7	81.2	2.0	91.9
16	209	2.0	86.1	2.9	91.4

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

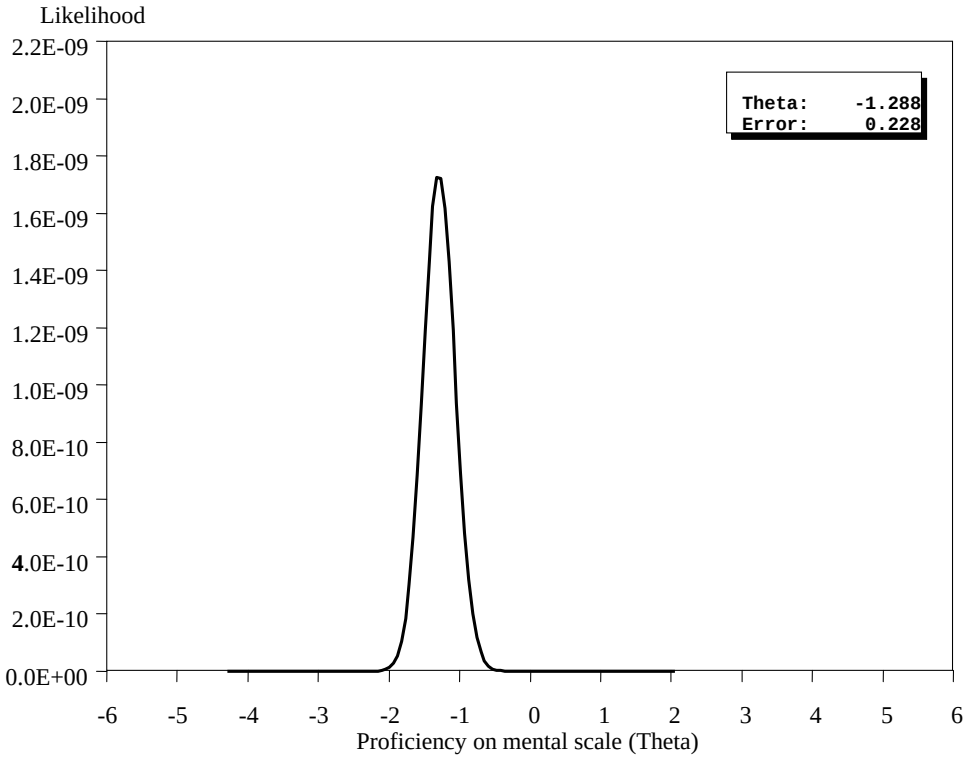
5.1 Bayley Short Form–Research Edition Scoring and Ability Estimates

In 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-PL logistic 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 one of the infants in publisher data is shown in figure 31, which shows that the likelihood of a response vector is quite small at any level of ability but appreciably smaller at some levels than others. Moreover, when the items and response vectors are informative, the range of more prominent likelihood values is constrained to a relatively short range. When the likelihood distribution is sharply concentrated, its graphical representation is similar to a spike. In this particular example, the infant 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 infant'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 infant represented by the figure would be $\overline{\theta}_i$ = −1.288.

Figure 31. 9-month BSF-R mental scale, response likelihood function for a specific examinee: Publisher data



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set of the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

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($\bar{\theta}_i$) = 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. For this reason, the maximum likelihood should not be used as if it were *the* ability estimate in a given instance. Instead, it is only the central tendency of the likelihood *distribution*.

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. This is also known as the Bayes estimate of the posterior distribution of θ, given response pattern $\mathbf{x}_i$. The EAP estimate is approximated using Gaussian quadrature, where $A(X_k)$ are normal ordinate weights for 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 of the IRT model but ignores errors that may result from equating to publisher scale metric.

5.4 Item Response Theory True Scores and Development Indices

The Psychological Corporation uses number-right scoring for the BSID-II mental and motor development scales. Raw scores are calculated by adding the number of the item immediately prior to the first item in the item set administered 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 publisher's regular 9-month item set, beginning with item number 63 was administered, 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.

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

[1] In the case of the 9-month BSF-R, the parametric model is a multilevel model that is age based. There is no single mathematical form of this model. This is done by age, so that each age has its own function. It cannot be reduced to a single form or formula.

5.5 ECLS-B Proficiency Level Probabilities

The BSF-R item response models provide interval scales along which every item and every examinee is positioned. The substantive significance of EAP ability estimates $\bar{\theta}_i$ 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 $\bar{\theta}_i$ 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 infant 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. 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. On this basis, the proficiency level probability scales presented in table 12 have been identified.

Table 12. Proficiency levels for Bayley Scales of Infant Development, Second Edition IRT 2-PL item calibrations using BSID-II standardization sample: 1993 and 2001–02

BSID-II Scale	Level	ECLS-B Subscale Proficiency Levels[1] BSF-R Subscale	Number of items	Normal deviate[1]	9-month percentile	Standard error
Mental	1	Explores objects in play	6	-3.56	0.3	0.50
	2	Explores purposefully	5	-2.20	18.9	0.46
	3	Babbles	4	-1.30	64.6	0.52
	4	Early problem solving	4	0.08	98.8	0.70
	5	Uses words	4	0.39	99.7	0.34
Motor	1	Demonstrates eye-hand coordination	4	-3.25	7.2	0.65
	2	Motor sitting	6	-3.09	9.8	0.43
	3	Motor pre-walking	6	-2.23	34.7	0.31
	4	Motor independent walking	5	-1.01	81.7	0.32
	5	Motor balance	4	0.57	99.5	0.45

[1] Normal deviate representing 67 percent of total credit possible on each sub-scale item set.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02; standardization data set of the Bayley Scales of Infant Development, Second Edition (BSID-II), The Psychological Corporation, 1993.

Subscales were constructed with publisher item calibrations by selecting the appropriate subset of items. By using the same item calibrations, the subscale score metric remains identical to that used in the corresponding main scale. Subscales vary in length from 4 to 6 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 a subscale with 4 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 range of subscale proficiency levels is broad, extending from the 1st to the 99th percentiles in terms of the mental development that is expected of 9-month infants, and from the 7th through the 67th percentile in terms of motor development. The selection of performance level subscales

is limited by the availability of items in the corresponding main scale. For this reason, it is not possible to define performance milestones at equal scale intervals.

In the case of the mental performance, Level 1 (Explores objects in play) represents an extremely low level of development for 9-month infants. For all practical intents and purposes, Level 1 can only be used to identify infants with severely deficient development. At the other extreme, Levels 4 (Early problem solving) and 5 (Uses words) identify infants that are highly developed. This leaves Levels 2 (Explores purposefully) and 3 (Babbles) as milestone events that are more appropriate for 9-month infants.

For the motor scale, Levels 1 (Demonstrates eye-hand coordination) and 2 (Motor sitting) are found at an extremely low level of development. Level 5 (Motor balance) is beyond the reach of most 9-month infants. In the middle range of motor development, Levels 3 (Motor pre-walking) and 4 (Motor independent walking) are more appropriate for 9-month infants.

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 infant can successfully execute each of the tasks belonging to the scale. By averaging over tasks, it is possible to calculate the probability of mastering a developmental milestone. Each cluster 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. The items follow 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.

Probabilities are calculated from the IRT true score after dividing by the total number of items in the subscale. When the resulting probabilities are plotted opposite EAP ability estimates, they represent the ICC for 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 are listed in table 13, 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. When the full BSID-II was intended to be included, it was necessary to obtain a measure of the child's age at the time of assessment, adjusted for prematurity. This is because the BSID-II Mental and Psychomotor Index scores are all based on the child's age at assessment, corrected for prematurity. This variable was programmed into the 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, however, are based on chronological age, not BSID-age. In addition, index scores were not obtained for the ECLS-B sample. There is no compelling reason to use BSID age, therefore. It was retained on the file for those few cases that were missing prematurity status from the birth certificate. This information was obtained from the parent respondent at the time of the child assessments. Therefore, this variable is the only place to find that information. It should not be used in analyses using the BSF-R scores, however.

Table 13. BSF-R mental scale and motor scale variable names, variable labels and range of values, 9-month data collection: 2001–02

Variable	Label	Range of values
X1MTLTSC	ECLS-B mental age-normed T-scores in N(50, 10) metric.	-41.83 to 117.33
X1MTLSCL	Mental IRT true score in publisher metric.	51.04 to 112.78
X1MTLSSE	Standard error of mental IRT true score.	2.35 to 8.19
X1MTL1	Explores objects in play	0-1
X1MTL2	Explores purposefully	0-1
X1MTL3	Babbles	0-1
X1MTL4	Early problem solving	0-1
X1MTL5	Uses words	0-1
X1MTRTSC	ECLS-B motor age-normed T-scores in N(50, 10) metric.	-33.39 to 98.46
X1MTRSCL	Motor IRT true score in publisher metric.	29.25 to 81.45
X1MTRSSE	Standard error of motor IRT true score.	1.53 to 7.55
X1MTR1	Demonstrates eye-hand coordination	0-1
X1MTR2	Motor sitting	0-1
X1MTR3	Motor pre-walking	0-1
X1MTR4	Motor independent walking	0-1
X1MTR5	Motor balance	0-1

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

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

The following tables, table 14 and table 15, summarize the average scores for the main grouping variables on the BSF-R mental and motor T-scores and scale scores, as well as the 5 mental probability scores and the 5 motor probability scores. With the exception of the one grouping variable of age at assessment, the means presented in these tables represent children of all ages, within each grouping variable, in the current data collection. 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 9-month BSF-R mental scale T-score (X1MTLTSC) was 48.59 for children living below poverty and 50.42 for children living at or above poverty level.

Table 14. Average BSF-R mental scale and mental probability scores by key demographic variables, 9-month data collection: 2001–02

Characteristic	Number	BSF-R mental scale mean scores (and standard deviation)						
		X1MTLTSC	X1MTLSCL	X1MTL1	X1MTL2	X1MTL3	X1MTL4	X1MTL5
Child race/ethnicity								
White	4,249	50.59 (9.69)	77.51 (7.57)	.99 (.01)	.91(.11)	.55 (.19)	.09 (.15)	.03 (.10)
Black	1,646	49.63 (10.04)	77.06 (7.60)	.99 (.02)	.90 (.13)	.54 (.19)	.09 (.14)	.03 (.09)
Hispanic, race specified	1,436	49.66 (10.33)	77.33 (7.40)	.99 (.01)	.91 (.12)	.55 (.19)	.09 (.14)	.03 (.09)
Hispanic, no race specified	645	48.19 (10.75)	76.73 (7.71)	.99 (.01)	.89 (.14)	.53 (.20)	.09 (.15)	.03 (.09)
Asian, non-Hispanic	1,112	48.47 (10.19)	76.89 (7.15)	.99 (.01)	.90 (.12)	.54 (.19)	.08 (.13)	.03 (.07)
Native Hawaiian, Pacific Islander	45	47.79 (8.12)	75.16 (6.01)	.99 (.01)	.88 (.12)	.49 (.16)	.05 (.10)	.01 (.04
American Indian, Alaska Native	274	47.44 (11.08)	79.77 (8.89)	79.77 (.01)	.92 (.11)	.61 (.22)	.14 (.19)	.05 (.13)
More than 1 race, Non-Hispanic	750	50.30 (9.98)	77.04 (7.24)	.99 (.01)	.90 (.12)	.54 (.19)	.08 (.14)	.03 (.09)
Poverty status								
Below poverty threshold	2,486	48.59 (10.45)	76.85 (7.87)	.99 (.02)	9 (.13)	.53 (.19)	.09 (.15)	.03 (.11)
At or above poverty threshold	7,709	50.42 (9.82)	77.46 (7.44)	.99 (.01)	.91(.12)	.55 (.19)	.09 (.15)	.03 (.09)
Sex								
Male	5,204	49.38 (9.98)	77.08 (7.49)	.99 (.01)	.90 (.12)	.54 (.19)	.09 (.14)	.03 (.09)
Female	4,991	50.64 (9.98)	77.58 (7.59)	.99 (.01)	.91 (.12)	.55 (.19)	.09 (.15)	.03 (.10)

See notes at end of table.

Table 14. Average BSF-R mental scale and mental probability scores by key demographic variables, 9-month data collection: 2001–02—Continued

Characteristic	Number	BSF-R mental scale mean scores (and standard deviation)						
		X1MTLTSC	X1MTLSCL	X1MTL1	X1MTL2	X1MTL3	X1MTL4	X1MTL5
Birth weight								
Normal (2,500 grams or more)	7,492	49.98 (9.92)	77.48 (7.53)	.99 (.01)	.91 (.12)	.55 (.19)	.09 (.15)	.03 (.10)
Moderately low (≥ 1,500 and < 2,500 grams	1,583	49.92 (10.66)	75.82 (7.43)	.99 (.01)	.88 (.15)	.51 (.19)	.07 (.14)	.03 (.09)
Very low (less than 1,500 grams)	1,082	51.73 (12.14)	73.44 (7.44)	.98 (.01)	.81 (.23)	.45 (.18)	.05 (.09)	.01 (.05)
Child age at assessment								
8 months or less	367	2.71 (8.15)	70.09 (3.97)	.98 (.02)	.76 (.17)	.36 (.09)	.01 (.01)	.00 (.00)
9–10 months	5,890	51.68 (8.14)	73.93 (4.42)	.99 (.01)	.87 (.12)	.46 (.12)	.03 (.03)	.00 (.01)
11–12 months	2,627	49.27 (10.15)	80.21 (5.44)	1.00 (.01)	.96 (.07)	.64 (.15)	.11 (.11)	.03 (.04)
13 months or more	1,311	43.00 (13.78)	89.48 (7.36)	1.00 (.00)	.99 (.02)	.85 (.13)	.35 (.23)	.17 (.20)
Maternal age (years)								
Less than 20	748	49.86 (10.48)	76.59 (7.24)	.99 (.01)	.90 (.12)	.53 (.18)	.08 (.14)	.03 (.09)
20–29	4,912	50.09 (10.09)	77.56 (7.67)	.99 (.02)	.91 (.12)	.55 (.19)	.09 (.15)	.03 (.10)
30–39	4,161	49.97 (9.73)	77.22 (7.43)	.99 (.01)	.91 (.12)	.54 (.19)	.09 (.14)	.03 (.09)
40+	303	49.73 (9.58)	75.82 (7.16)	.99 (.01)	.88 (.14)	.51 (.19)	.07 (.12)	.02 (.06)

See notes at end of table.

Table 14. Average BSF-R mental scale and mental probability scores by key demographic variables, 9-month data collection: 2001–02—Continued

Characteristic	Number	BSF-R mental scale mean scores (and standard deviation)						
		X1MTLTSC	X1MTLSCL	X1MTL1	X1MTL2	X1MTL3	X1MTL4	X1MTL5
Mother's education								
8th grade or less	520	46.79 (10.21)	76.35 (7.54)	.99 (.02)	.88 (.14)	.52 (.19)	.08 (.14)	.03 (.09)
9–12th grade	2,157	49.50 (9.83)	77.22 (7.81)	.99 (.01)	.90 (.13)	.54 (.20)	.09 (.16)	.03 (.10)
High school diploma	2,156	49.65 (10.22)	77.26 (7.49)	.99 (.01)	.90 (.12)	.55 (.19)	.09 (.14)	.03 (.09)
Voc/technical	212	51.45 (10.39)	77.16 (6.04)	.99 (.01)	.92 (.11)	.55 (.17)	.07 (.10)	.02 (.04)
Some college	2,400	50.33 (9.94)	77.86 (7.79)	.99 (.01)	.91 (.10)	.56 (.20)	.10 (.16)	.04 (.11)
Bachelor's degree	1,606	51.02 (9.32	77.12 (7.15)	.99 (.01)	.91 (.10)	.54 (.19)	.09 (.14)	.03 (.08)
Grad. school (no degree)	177	51.64 (9.13)	76.33 (6.63)	.99 (.01)	.90 (.12)	.52 (.17)	.07 (.11)	.02 (.06)
Master's degree	651	51.12 (10.56)	77.20 (7.16)	.99 (.01)	.92 (.12)	.55 (.19)	.09 (.13)	.03 (.06)
Doctoral/prof. deg	242	52.27 (8.75)	78.41 (8.19)	.99 (.01)	.92 (.12)	.57. (.20)	.11 (.18)	.05 (.11)

NOTE: Results were obtained by using the sampling child weight W1C0.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

Table 15. Average BSF-R motor scale and motor probability scores by key demographic variables, 9-month data collection: 2001–02

Characteristic	Number	BSF-R motor scale mean scores (and standard deviation)						
		X1MTRTSC	X1MTRSCL	X1MTR1	X1MTR2	X1MTR3	X1MTR4	X1MTR5
Child race/ethnicity								
White	4,240	49.67 (10.00)	56.23 (6.63)	.92 (.04)	.95 (.06)	.78 (.20)	.29 (.30)	.06 (.15)
Black	1,639	52.49 (9.81)	57.56 (6.50)	.93 (.07)	.96(.07)	.82 (.17)	.36 (.31)	.07 (.15)
Hispanic, race specified	1,432	48.81 (9.97)	56.06 (6.60)	.92 (.07)	.94 (.07)	.77 (.20)	.29 (.29)	.05 (.14)
Hispanic, no race specified	643	49.50 (9.90)	56.39 (6.80)	.92 (.08)	.94 (.07)	.78 (.21)	.31 (.31)	.06 (.14)
Asian, Non-Hispanic	1,106	50.00 (9.83)	56.75 (6.62)	.93 (.07)	.95 (.07)	.79 (.19)	.32 (.31)	.06 (.14)
Native Hawaiian, Pacific Islander	45	55.28 (7.09)	58.14 (5.34)	.95 (.03)	.97 (.03)	.86 (.12)	.37 (.29)	.06 (.14)
American Indian, Alaska Native	274	50.43 (11.10)	59.02 (7.91)	.94 (.06)	.96 (.06)	.83 (.19)	.42 (.35)	.12 (.22)
More than 1 race, Non-Hispanic	749	50.20 (10.31)	57.02 (6.50)	.93 (.07)	.95 (.06)	.81 (.19)	.34 (.30)	06 (.14)
Poverty status								
Below poverty threshold	2,482	48.59 (10.45)	56.69 (6.74)	93 (.07)	95 (.07)	.79 (.20)	.32(.31)	.06 (.15)
At or above poverty threshold	7,684	49.94 (9.91))	56.40 (6.62)	.92 (.07)	.95 (.07)	.78 (.20)	.30 (.30)	.06 (.15)
Sex								
Male	5,185	49.94 (9.93)	56.48(6.60)	.92(.07)	.95(.07)	.79(.19)	.31(.30)	.06(.15)
Female	4,981	50.07 (10.07)	56.45(6.69)	.92(.07)	.95(.07)	.78(.20)	.31(.30)	.06(.15)

See notes at end of table.

Table 15. Average BSF-R motor scale and motor probability scores by key demographic variables, 9-month data collection: 2001–02—Continued

Characteristic	Number	BSF-R motor scale mean scores (and standard deviation)						
		X1MTRTSC	X1MTRSCL	X1MTR1	X1MTR2	X1MTR3	X1MTR4	X1MTR5
Birth weight								
Normal (2,500 grams or more)	7,483	50.17(9.83)	56.67(6.58)	.93(.06)	.95(.06)	.79(.19)	.31(.30)	.06(.15)
Moderately low (≥ 1,500 and < 2,500 grams	1,574	48.20(11.43)	54.61(6.70)	.90(.09)	.93(.09)	.72(.23)	.24(.27)	.04(.12)
Very low (less than 1,500 grams)	1,071	46.24(12.79)	50.89(7.04)	.84(.14)	.86(.15)	.59(.27)	.14(.20)	.02(.06)
Child age at assessment								
8 mos. or less	366	50.55(9.34)	50.70(4.90)	.86(.09)	.88(.11)	.59(.22)	.09(.10)	.00(.01)
9-10 mos.	5,876	50.57(8.46)	53.78(4.33)	.91(.06)	.93(.07)	.72(.19)	.17(.16)	.01(.02)
11-12 mos.	2,617	47.93(10.72)	58.40(5.06)	.95(.05)	.97(.04)	.87(.14)	.40(.28)	.05(.09)
13 mos. or more	1,307	50.92(13.95)	66.77(6.44)	.98(.31)	.99(.02)	.97(.07)	.80(.25)	.30(.28)
Maternal age (years)								
Less than 20	746	51.93(9.56)	56.91(6.16)	.93(.06)	.96(.06)	.81(.18)	.32(.30)	.05(.13)
20-29	4,902	50.52(9.79)	56.81(6.68)	.93(.07)	.95(.07)	.80(.19)	.32(.30)	.06(.16)
30-39	4,144	49.18(10.21)	56.07(6.68)	.92(.07)	.94(.07)	77(.20)	.29(.30)	.05(.14)
40 +	303	47.70(9.92)	54.50(6.06)	.90(.08)	.93(.08)	.73(.21)	.23(.27)	.03(.08)

See notes at end of table.

Table 15. Average BSF-R motor scale and motor probability scores by key demographic variables, 9-month data collection: 2001–02—Continued

Characteristic	Number	BSF-R motor scale mean scores (and standard deviation)						
		X1MTRTSC	X1MTRSCL	X1MTR1	X1MTR2	X1MTR3	X1MTR4	X1MTR5
Mother's education								
8th grade or less	520	48.50 (10.88)	56.17 (7.06)	.92 (.08)	.94 (.08)	.77 (.21)	.30 (.30)	.06 (.16)
9–12th grades	2,150	50.67 (9.96)	56.87 (6.64)	.93 (.07)	.95 (.07)	.80 (.19)	.33 (.31)	.06 (.15)
High school diploma	2,156	50.45 (10.18)	56.71 (6.72)	.93 (.07)	.95 (.07)	.79 (.20)	.32 (.31)	.06 (.15)
Voc/technical	212	48.03 (9.37)	55.13 (5.29)	.92 (.06)	.94 (.06)	.76 (.19)	.24 (.24)	.03 (.07)
Some college	2,392	50.34 (9.74)	56.86 (6.72)	.93 (.06)	.95 (.06)	.80 (.19)	.32 (.31)	.06 (.16)
Bachelor's degree	1,603	49.69 (9.68)	55.84(6.53)	.92 (.06)	.94 (.07)	.77 (.20)	.27 (.29)	.05 (.15)
Grad. school (no degree)	177	48.11 (10.43)	54.49 (5.63)	.91 (.07)	.94 (.07)	.74 (.19)	.20 (.22)	.03 (.12)
Master's degree	650	48.41 (10.02)	55.48 (6.52)	.91 (.07)	.94 (.07)	.75 (.21)	.27 (.29)	.05 (.13)
Doctoral/prof. deg	241	50.01 (8.99)	56.45 (5.89)	.93 (.06)	.95 (.06)	.80 (.20)	.32 (.29)	.04 (.08)

NOTE: Results were obtained by applying the sampling child weight W1C0.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

This page is intentionally left blank.

6. NURSING CHILD ASSESSMENT TEACHING SCALE IN THE ECLS-B

To capture the full breadth of infant functioning, it was important to include a direct measure of young children's socioemotional functioning. The following section outlines the rationale for obtaining an observational measure and the decision process that led to the selection of the most appropriate measure for the ECLS-B, the Nursing Child Assessment Teaching Scale (NCATS). This is followed by a description of the in-home administration of the NCATS as well as the quality control procedures that were undertaken to ensure that the data obtained were of high quality, including interviewer training, training the trainers on the coding system, and training the coders. In addition, a summary of coder reliability and Cronbach's alpha for the NCATS subscales are presented followed by a summary of how these measures performed in the national data collection.

During infancy, socioemotional functioning is easiest to obtain during mother-child interaction because it provides a context in which infant emotional functioning can be elicited and observed. At this age, infants have begun to recognize and to produce emotional expressions using the communicative cues of close caregiving adults (Tronick 1989). This is supported by multiple lines of developmental research in such areas as intersubjectivity (Trevarthen and Aitken 2001); social referencing (Walker-Andrews 1998); attunement (Stern 1985); and emotion regulation (Miller, McDonough, Rosenblum, and Sameroff 2002). Many of the important socioemotional behaviors that develop during infancy are difficult to measure in the child independently because the important processes contributing to this developmental progress are most clearly observed in interactions between parent and child. For example, constructs such as temperament, attention, emotion and state regulation, communication, cognition, and even some areas of motor development are mediated by interactions with primary caregivers. Therefore, the Technical Review Panel members considered it important to observe mother-child interaction during a standardized situation in order to evaluate infant socioemotional functioning directly.

Several measures were considered, but the field quickly narrowed to the NCATS on the advice of the Technical Review Panel members. The following sections provide overviews of the steps that went into the selection of the NCATS, its implementation in a home context, the trainings that were required to ensure adequate videotape quality and reliable coding, and the steps that were taken to ensure inter-lab reliability with the developer of the NCATS. (To obtain further information about the NCATS,

or to place an order for a manual, the analyst is referred to the Nursing Child Assessment Satellite Training (NCAST) web site: http://www.NCAST.org.)

6.1 Technical Review Panel Advice

Technical Review Panel members recommended the NCATS for the 9-month data collection of the ECLS-B because it is relatively easy to administer, is psychometrically sound, and correlates with children's outcomes, family backgrounds, and the quality of the home environment. The inclusion of the NCATS in the ECLS-B 9-month data collection was sponsored by the Administration for Children, Youth and Families (ACYF) of the U.S. Department of Health and Human Services.

The NCATS has been used with success in several large scale studies including the Early Head Start Research and Evaluation Project (with a national sample of approximately 3,000 infants) sponsored by ACYF, the Early Intervention Collaborative Study (EICS; Shonkoff, et al. 1992), the National Institute of Child Health and Human Development (NICHD) Early Child Care Study, the Evaluation of the Comprehensive Child Development Program (CCDP), and the Memphis New Mothers Study.

The use of the NCATS is well supported by the Nursing Child Assessment Satellite Training (NCAST), publisher of the NCATS, an important consideration in deciding to include the NCATS in the 9-month ECLS-B. This support includes standardized trainings for trainers, standardized educational and training materials, and maintenance of inter-rater reliability. This level of support was seen as a significant advantage that would help ensure the collection of high quality data that would be comparable to data obtained in other studies, such as the Early Head Start Research and Evaluation Project. For the purposes of the ECLS-B 9-month data collection, the NCATS was seen by Technical Review Panel members as the most viable measure of parent-child interaction because it is one of the few field-tested systems with excellent training materials, good psychometric properties, and, while brief, produces scores predictive of later growth in both cognitive and social-emotional domains.

The NCATS is a part of a larger parent-child observation and intervention system called the Nursing Child Assessment Satellite Training (NCAST; Barnard, 1978). The NCAST package consists of two activities, a feeding task and a teaching task, for observing mother and child. There are six subscales scored for each task, as well as a set of assessment and early intervention programs that train service

providers in clinical settings to assess and intervene to improve parent-child interactions. In the observation component, the two tasks are the Nursing Child Assessment Feeding Scale (NCAFS) and (2) The Nursing Child Assessment Teaching Scale (NCATS). Thus NCATS refers to the specific observational system for administering and scoring the teaching task while NCAST refers to the entire assessment and intervention package.

The NCATS assesses characteristics of the parent-child relationship from birth through age 3 that are known to be important predictors of social and cognitive development when children are 2 and 3 years of age and older, such as parent's growth fostering and the child's responsiveness to the parent. (Sumner and Spietz, 1994: Barnard, 1997). Developed by Dr. Kathryn Barnard at the University of Washington School of Nursing, the NCATS was designed to provide information about the nature of parent-child interactions for both research and clinical intervention. In the NCATS, the parent is asked to teach the child a task, such as stacking blocks, that is slightly beyond the child's current abilities and therefore creates some stress for the child. The focus of the NCATS is the interaction between the parent and the child rather than the child's success or failure at learning the task. A particular focus is the mother's responsiveness to the child's distress and the child's responsiveness to the mother.

6.2 NCATS ECLS-B Nine-Month Protocol for In-Home Administration

It was not operationally feasible to train interviewers to code the NCATS live during the home visit due to all the other tasks that they were required to perform. Therefore, interviewers administering the NCATS during the home visit used a handheld video camera to videotape the parent and child engaging in the NCATS teaching task. During the national training, interviewers were taught to administer and to videotape the NCATS. The training included practice emphasizing good filming techniques and skillful use of the camera in conjunction with faithful administration of the NCATS task. An emphasis was placed on creating videotapes of the same quality as the NCATS practice videotapes provided by NCAST at the University of Washington School of Nursing. In addition, the *Field Representative Manual* was provided to all trainees and included detailed instructions on videotaping and administering the NCATS. Interviewers were able to refer to this manual during the field period as needed.

NCATS administration during the home visit was standardized to ensure that all interviewers administered the NCATS the same way to all parent-child pairs. To ensure this standardization, step-by-

step NCATS administration instructions were included in the Child Activity Booklet in a separate tabbed section for NCATS. These instructions also included a verbatim script that was read to the parent. Interviewers were instructed to obtain the NCATS in all cases, regardless of the languages used by the mother to the child. Interviewers then checked which one of the 15 possible NCATS activities the mother chose to teach to the child. In addition, while following the instructions, interviewers also recorded key information where required, for example, the activity chosen by the parent, the start time of the NCATS, and the language used by the parent to teach the child.

In the case of twins, the interviewer administered a separate NCATS for each twin, each on separate videotapes and using different activities. In addition, to assess maternal and child behaviors in the context of the triad (that is, the mother and twins simultaneously), a third NCATS videotape was obtained of the mother teaching yet a third activity to both twins in a triadic interaction. However, these triadic videotapes have not yet been coded pending modifications to the NCATS coding system, which does not accommodate triadic interactions.

After completion of the home visit, the field representative then sent the NCATS videotape (or videotapes in the case of twins) and the Child Activity Booklet, along with other data collection materials, to Westat's home office for receipting and coding. The interaction could be coded "live" by a trained coder, or from videotape of the interaction. As noted earlier, for the ECLS-B, field representatives videotaped the parent-child interaction. The tapes were then sent to Westat to be coded by trained coders at the home office. Westat had bilingual Spanish coders to code the Spanish tapes. In addition, because Westat has a diverse staff, videotapes in other less frequently encountered languages were also coded by recruiting available staff members. All NCATS items that could be coded independent of language were coded first, with the language-dependent items reserved until an in-house interpreter was identified. In this way, Westat was able to obtain complete NCATS codes on all but approximately 48 videotapes of interactions that took place in such languages as Daria, Dinka and, Laotian.

6.3 Field Testing of NCATS Coding Procedures

Quality control and coding procedures for the NCATS were tested over a 2-year period and two evaluation studies were conducted. The first evaluation study, conducted on a randomly selected subsample of 150 videotapes, assessed coder reliability in both labs (i.e., the University of Washington's certified coder and Westat's trained coders) and within Westat (i.e., Westat's coders with the Westat

coding supervisor). As a result of this testing, training procedures for coders and quality control standards to maintain coder reliability were designed and implemented. The second evaluation study served two purposes. The first purpose was to assess Westat's ability to establish a coding staff with the capacity to code reliably approximately 800–1000 videotapes per month. The second purpose was to evaluate the success of improvements to training and to coding procedures that were developed in response to the findings from the first evaluation study. These training procedures and certification procedures and their results are presented in the following sections.

6.4 Three NCATS Trainings

Comprehensive training was required to make sure that videotapes were of sufficiently high quality to be codable and to ensure that coding was done reliably. Hence, three different types of training were required. First, field staff were trained to obtain high quality videotapes and to administer the NCATS to the parent and child according to standardized procedures. Second, home office staff targeted to train coders on the NCATS coding system attended a standardized training session at the University of Washington. Third, NCATS coders participated in in-depth training to ensure reliability of coding comparable to University of Washington standards. Each of these trainings is described in more detail in the following sections. This is 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 presents reliability coefficient alphas for the NCATS subscales, comparing Westat coders with the coders at the University of Washington.

6.4.1 Field Staff Training

Attention was directed to two areas during the 9-month national NCATS training of field staff. The first concentration was on correct administration of the NCATS procedures. Here, field staff were trained to follow the NCATS administration steps verbatim as presented in the Child Activity Booklet. They then administered the NCATS to each other in sets of three trainees, in alternating turns, one to play the role of the interviewer, one the parent, and one the child. The second concentration was on proper videotaping practices and use of the camera. Obtaining a high quality videotape was critical to successful NCATS coding. Therefore, field staff trainees received hands-on practice and detailed feedback about their videotaping. This was done during the training sessions 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 in their practice sessions. In addition, field staff videotaped each other during the live practice session while the partner administered the BSF-R. When Westat staff members reviewed the videotape to score the BSF-R administration, the quality of the videotaping was evaluated. Any videotape that was not of sufficient quality (e.g., audio level too low, lighting level too low, camera faced toward a window so the dyad was seen only in silhouette, camera positioned too closely to a source of a loud noise so speech was inaudible) 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, the first videotapes for each field staff member that were received at Westat were quality reviewed by the NCATS coding staff immediately upon receipt. Feedback was given about videotape quality to all field staff within about a week of receipt of the first videotapes. For further information about quality control procedures, please see section 6.5.

6.4.2 Trainer Training

The first step was to have the trainers 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 NCAST at the University of Washington. As mentioned, the NCATS can be a "turn-key" package that includes a standardized training for the intended trainers. It is a requirement of NCAST that individuals wishing to train others to code the NCATS must attend the trainer training at the University of Washington. The reason for this is that the training for trainers is more rigorous than the training for those who will code the NCATS but not teach it. NCAST implemented these higher standards in order to minimize error in coding by preventing attenuation of standard practice as these trainers passed the information on to coder trainees.

In the summer of 2001, four Westat staff members from the Child and Family Studies area attended the standardized training at the NCAST center at the University of Washington. The training took place over the span of 5 days and was conducted by a certified NCAST trainer. To pass the training, the individual had to view five reliability videotapes and code each with a score of 90 percent or higher for each of the subscales. The subscales include the four caregiver subscales—caregiver sensitivity to cues; caregiver response to child's distress; caregiver social-emotional growth fostering; and caregiver cognitive growth fostering—and the two child subscales, clarity of cues and responsiveness to caregiver.

NCAST allows an individual up to three attempts to pass the training with a score of 90 percent or better on each of the subscales. Two of Westat's designated trainers passed all five videotapes with a score of 90 percent or greater for each of the subscales on the first try. An additional staff member had to redo one videotape, but passed the other four on the first try. Finally, one staff member had to redo two videotapes once, but passed the other three on the first try. The end result was that all four staff members completed the NCAST training with a score of 90 percent or higher on all six subscales for all five reliability training tapes.

One staff member, because of her involvement as lead coding trainer during the field tests, led the first training of coders that took place shortly after completion of the NCAST training. A second individual co-led this training and was designated to be the supervisor of the Westat coders. Because her leadership would shape the interpretations of coding issues, it was important that she be trained to the higher trainer standards. A third individual, a member of the instrument development team, served as the liaison between the Child and Family Studies area and the NCATS coding workshop. This individual, because of her training as a trainer, was also called upon to resolve coding questions that arose during the course of the year. The coding supervisor and the liaison were able to establish, maintain, and share with coders an NCATS coding knowledge base that contributed to the maintenance of coding reliability across the data collection year. Because videotape coding is tedious and demanding, it was anticipated that there would be considerable attrition in coding staff, which would require repeated coder trainings. To reduce the burden of training to the coding supervisor and to the Child and Family Studies area liaison, the additional individual trained by NCAST would serve as lead trainer for any trainings made necessary by attrition. As it turned out, there was no attrition and further trainings did not take place, although this could not have been known in advance.

6.4.3 Coder Training

Recruitment of potential coders through local temporary employment agencies (supplemented by Westat hourly personnel) began immediately after the trainer training. A screening process was developed to maximize the efficiency of recruitment and to target individuals with good observational skills. Potential candidates were first asked to complete an observation skills test that had been developed through pilot testing at Westat. This observation skills test presented either a detailed line drawing or a detailed photograph of a common street scene. The candidate was instructed to view each picture for 5 minutes. Then the candidate was instructed to turn the page (with no turning back) and

answer approximately 8-10 multiple choice questions about the details in the scene. There were six sets of pictures and a total of 50 questions (approximately 8 questions per picture).

Pilot testing at Westat, involving individuals who already been trained on NCATS coding for the field test, had determined that 90 percent accuracy was the minimum cut-off to ensure good observation skills; those scoring 90 percent or higher on the observation skills test had NCATS reliabilities of 85 percent agreement or higher on their weekly NCATS reliability tapes during the field test. The observation skills test was administered to approximately 60 individuals. Of these, 28 individuals passed the observation skills test at 90 percent. These 28 individuals then participated in three waves of self-paced tutorials.

The self-paced tutorial had been suggested by Dr. Barnard as a way of familiarizing trainees with the coding system as well as with the demands of working as a coder. For this tutorial, each candidate received a copy of the NCATS manual (Sumner and Spietz 1994) and a sample NCATS score sheet to review. They also viewed training videos provided by NCAST, including videotapes about children's engagement and disengagement cues and background information about the Barnard model, as well as the coding instructions. Candidate trainees then completed the reliability pre-screening in which they coded five videotaped interactions obtained during the field test. The required passing score on this reliability pre-screening was 85 percent.

Trainees who passed the self-paced tutorial at 85 percent agreement or higher were then interviewed. During this interview, conducted by a Child and Family Studies staff member, the demands of observational coding were explained and the candidate's suitability for and commitment to employment as a coder were assessed. As a result of the 85 percent agreement criterion during the self-paced tutorial and the screening interview, there were 18 individuals who were advanced to the formal training. All of these individuals then went on to pass the 90 percent agreement criterion for the formal training.

Coder training was held in three waves beginning at the end of October 2001 and took 4 days per wave. The same two lead trainers conducted the initial three waves of trainings. Training consisted of 2 days of lectures to review the content and intention of the items combined with practice coding of the items after viewing practice videotapes. The third day was devoted to reliability testing to certify coders. To be certified, trainees had to code five videotaped parent-child interactions and reach the required 90 percent agreement with NCAST reliability coding. The actual scoring of the reliability testing

for certification was done by certified NCAST staff at the University of Washington. While trainees waited for reliability feedback from NCAST, they practiced coding field test videotapes. As soon as a trainee was certified, she or he began coding NCATS videotapes from the national data collection. Trainees who scored less than 90 percent agreement were then re-tested on either an entire videotape or only on specific subscales per videotape, as necessary. Ultimately, all 18 trainees passed the reliability testing with 90 percent or higher for each subscale within the allowable three attempts.

Coding of NCATS tapes then began as coders became certified. In keeping with the recommendation of Dr. Barnard, coders worked up to 4 hours a day, to a maximum of eight videotapes. It was Dr. Barnard's opinion that coding reliability begins to falter beyond that amount. Initially, all 18 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, receipting, computer-assisted data entry (CADE), and so forth. However, as the sample size of the ECLS-B was reduced, about halfway through the data collection, fewer videotapes were expected each month. Therefore, a number of coders were released from their coding duties in two waves. For each wave of coder reductions, the criteria for retaining coders included high reliabilities and coding productivity as the most important criteria, followed by usability in other departments and finally whether the coder was a Westat hourly employee. By the middle of the data collection, there were eight coders, all of whom had demonstrated good reliability and productivity.

6.5 NCATS Coding Quality Control Procedures

There were two important aspects of quality control that were critical to the success of the NCATS in the ECLS-B. The first was the maintenance of inter-lab reliability between the Westat coders and the certified NCATS coders at the University of Washington. In addition to requiring that coders be initially certified to the 90 percent reliability standards established by Dr. Barnard's lab at the University of Washington School of Nursing, Westat also tested and monitored the interlab inter-rater reliability of the coders on an ongoing basis. Use of a higher reliability standard in the training phase ensured that, even with the likely occurrence of some degree of coder "drift," ongoing reliability would still be well within acceptable ranges. Therefore, based on the NCATS manual and discussion with Dr. Barnard, the criterion for ongoing reliability was 85 percent agreement between each NCATS coder and the reliability coder at the University of Washington. Interlab inter-rater reliability was obtained only on English tapes. It was not possible to obtain reliabilities separately on Spanish tapes because the University of Washington did not have a certified Spanish-speaking coder.

Inter-lab reliability is a reflection of Westat coders' ability to maintain the high coding standards learned during training throughout the year of data collection without drifting from that standard. Therefore, a random subset of approximately 20 videotaped NCATS interactions per month were copied and sent to the certified NCATS reliability coder at the University of Washington for reliability coding. It was originally intended that videotapes totaling 2½ percent of the sample would be reliability coded by NCAST. However, only a total of 171 tapes were reliability coded. This number actually proved sufficient to provide all coders with the target of one reliability tape each per week. In addition, because the same coders who coded videotapes in Spanish also coded videotapes in English, it was not necessary to obtain separate reliabilities for English and Spanish videotapes.

Over the year of coding, the inter-lab reliability, as measured in percent agreement between each Westat coder and the certified NCAST coder, was satisfactory with the an overall average agreement between Westat and NCAST of 86 percent.

If a coder slipped below 85 percent agreement on a weekly reliability tape, that coder then immediately coded a second reliability tape. If the second reliability tape was also below 85 percent agreement, the coder was told to cease coding any tapes from the ECLS-B and the coding supervisor intervened in one of several ways, as appropriate:

- Discussion between the coder and supervisor to resolve disagreements about problem items.
- More extensive one-on-one retraining, provided by the coding supervisor, tailored to address any content areas that needed improvement.
- Additional NCATS coding practice using videotapes drawn from the ECLS-B field test, but not 9-month ECLS-B tapes.
- Review of the NCAST training videotapes supplied from the original training.

Coder reliability is an important but not the only measure of reliability. Coefficient alpha is an estimate of within-measure consistency, that is, how well the items in an instrument cohere with each other in measuring a construct. The value of alpha can range from 0 to 1. An alpha of 1 indicates that the items are all measuring the same construct. An alpha of 0 indicates that the items are inconsistent and may not be measuring the same construct. According to Nunnally (1978), a satisfactory level of reliability depends on how a measure is being used. For research purposes, a reliability of .70 or higher is sufficient

although alphas of at least .60 or greater are often found in research As noted earlier, there are four NCATS subscales that typify the caregiver's behavior and two that describe the child's behavior. The subscales describe the following: caregiver's sensitivity to cues; caregiver's response to the child's distress; caregiver's social-emotional growth fostering; caregiver's cognitive growth fostering; child's clarity of cues; and child's responsiveness to the caregiver (Sumner and Spietz 1994).

Table 16 compares the subscale and total scale alphas obtained by Westat and NCAST coders on the same subset of 171 randomly selected reliability tapes. The two middle columns of table 16 compare the alphas obtained by Westat coders and by the NCAST coders. The rightmost column presents alphas from the NCAST manual. These alphas were obtained on 1,887 cases that have been provided to NCAST by university researchers who have used the NCATS over the past couple of decades. Comparison of the alphas shows that the alphas published in the manual are higher than the alphas obtained in the ECLS-B 9-month data set, by both Westat and NCAST. Possible explanations for this include the videotaped format that was used for coding. It is possible that videotaping changes the quality of the mother-child interaction or changes the nature of the coding. Changes in coding could be due to differences in what is visible on video and what is visible in person. Coding differences could also be due to the ability to rewind a videotape to clarify whether one's perception is correct, e.g., was that really a smile? Alternatively, it is possible that differences in the research samples explain these differences in that the samples aggregated into the University of Washington data set are a convenience sample whereas the ECLS-B sample is nationally representative.

Table 16. Reliability correlation coefficients (standardized) for subscales of the NCATS Teaching Scale for ECLS-B 9-month data collection, NCAST coders, and NCAST Manual

NCATS scale	9-month ECLS-B (N=8,520)	NCAST coders (N=171)	University of Washington (N=1,887)
Sensitivity to cues	.12	.27	.52
Response to distress	.66	.52	.80
Social-emotional growth fostering	.34	.45	.58
Cognitive growth fostering	.58	.62	.78
Clarity of cues	.39	.36	.50
Responsiveness to caregiver	.58	.58	.78
Total Parent	.68	.74	.87
Total Child	.62	.63	.80
Overall Total Score	.72	.77	.87

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

That said, the 9-month ECLS-B NCATS data show that several of the subscales have low alphas. This suggests that these subscales do not measure unitary constructs. For that reason, the 9-month ECLS-B data file includes the item-level NCATS data, as well as the Total NCATS score, the Total Parent score, and the Total Child score, the latter three having acceptable alphas. This enables the analyst to examine the item content of the NCATS to determine whether other item groupings would be more suitable for the analyst's research needs.

Careful monitoring of videotape quality was the second important aspect of quality control. There were two quality control forms that coders filled out. The first was a videotape quality control checklist of whether the videotape was codable. If not, then the coder used the checklist to identify the reason(s) why the videotape was not codable, for example, poor lighting, poor audio level, interference by siblings/other caregiver(s), use of a non-NCATS toy (which invalidates the coding), foreign language, and blank tape. The other quality control form obtained information for in-house use about the specific language, the length of the interaction, the amount of time it took to code the videotape, and so forth. A fast feedback system was implemented in time for the beginning of data collection that sent out feedback to field supervisors about any videotaping problems an interviewer had on a home visit. This feedback was sent out within a week after receipt of the tape and the field supervisor was advised about what the problem was and what feedback to give the interviewer to improve videotape quality. This was done for

every videotape receipted, although feedback was sent to supervisors only if there were problems with a videotape.

6.6 NCATS Scale Scores in the 9-Month Data Collection

The 9-month data set includes all the item-level NCATS variables because the analyst may want to use factor analysis to explore the possibility that an alternate structure may obtain subscales that have higher alphas than those presented in table 16. Composites of the standard recommended total scale scores, however, are already available in the 9-month data set, as listed in table 17.

Table 17. NCATS Scale scores, variable names, variable labels and range of values, 9-month data collection: 2001–02

Variable	Label	Range of values
X1NCATTS	NCATS Total Scale score	0-73
X1NCATTP	NCATS Total Parent Scale score	0-50
X1NCATTC	NCATS Total Child Scale score	0-23

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

Table 18 summarizes the descriptive statistics of the three NCATS scores in the 9-month data collection for the sample as a whole and by the same demographic variables as in the above sections. The additional demographic of mother's race was added because the mother's behaviors during the interaction are represented in the NCATS Total Parent Scale score (NCATTP) and the NCATS Total Scale score. In addition, each cell includes the number of cases because there can be a child scale score without a parent scale score, and vice versa, resulting in some variability in the "n's." The number of cases can also vary if the value of the child weight, W1C0, equals zero. To obtain these statistics, reserve codes were deleted and the data were weighted using the child weight W1C0.

Table 18. Average NCATS Scale scores by demographic variables, 9-month data collection: 2001–02

Characteristic	NCATS Scales mean scores and standard deviations								
	Total Scale score (X1NCATTS)			Total Parent Scale score (X1NCATTP)			Total Child Scale score (X1NCATTC)		
	n	Mean	SD	n	Mean	SD	n	Mean	SD
Child race/ethnicity									
White	3,701	50.94	5.66	3,701	35.44	4.41	3,700	15.50	2.68
Black	1,365	49.47	5.84	1,364	33.92	4.44	1,365	15.55	2.73
Hispanic, race specified	1,172	49.22	5.80	1,172	33.65	4.56	1,171	15.56	2.75
Hispanic, no race specified	536	48.34	5.62	536	32.94	4.21	536	15.40	2.68
Asian, non-Hispanic	869	49.61	5.82	869	34.47	4.38	868	15.14	2.71
Native Hawaiian, Pacific Islander	34	49.54	8.33	34	34.71	6.27	34	14.83	2.93
American Indian, Alaska Native	244	48.79	5.78	244	33.74	3.94	244	15.04	3.03
More than 1 race, Non-Hispanic	656	50.49	5.30	656	34.86	4.24	656	15.63	2.65
Mother race/ethnicity									
White	4,057	50.97	5.67	4,057	35.45	4.41	4,056	15.52	2.68
Black	1,387	49.43	5.81	1,387	33.91	4.44	1,388	15.52	2.74
Hispanic, race specified	1,440	48.56	5.69	1,440	33.10	4.41	1,439	15.46	2.72
Hispanic, no race specified	16	48.42	3.52	16	32.28	2.07	16	16.14	2.21
Asian, non-Hispanic	1,027	49.76	5.71	1,027	34.58	4.28	1,026	15.17	2.72
Native Hawaiian, Pacific Islander	41	51.05	5.05	41	35.12	3.92	41	15.93	2.47
American Indian, Alaska Native	336	48.96	5.56	336	33.78	4.01	336	15.17	2.76
More than 1 race, Non-Hispanic	242	50.96	5.17	242	35.10	4.05	242	15.86	2.72
Poverty status									
Below poverty threshold	2,074	48.54	5.72	2,074	33.17	4.39	2,073	15.36	2.76
At or above poverty threshold	6,534	50.66	5.70	6,534	35.11	4.45	6,533	15.54	2.68
Sex									
Male	4,388	40.01	5.86	4,388	34.57	4.56	4,387	15.44	2.69
Female	4,220	50.37	5.67	4,220	34.79	4.45	4,219	15.57	2.71

See notes at end of table.

Table 18. Average NCATS scale scores by demographic variables, 9-month data collection: 2001–02—Continued

Characteristic	NCATS Scales mean scores and standard deviations								
	Total Scale score (X1NCATTS)			Total Parent Scale score (X1NCATTP)			Total Child Scale score (X1NCATTC)		
	n	Mean	SD	n	Mean	SD	n	Mean	SD
Birth weight									
Normal (2,500 grams or more)	6,353	50.25	5.76	6,353	34.72	4.51	6,350	15.53	2.69
Moderately low (≥ 1,500 and < 2,500 grams	1,323	49.52	5.83	1,323	34.22	4.41	1,323	15.29	2.77
Very low (less than 1,500 grams)	899	48.60	5.77	899	33.77	4.63	900	14.83	2.71
Child age at assessment									
8 mos. or less	312	49.82	5.96	312	34.44	4.69	312	15.38	2.61
9–10 mos.	5,054	49.81	5.90	5,054	34.39	4.62	5,051	15.42	2.70
11–12 mos.	2,173	50.60	5.67	2,173	34.89	4.32	2,173	15.71	2.74
13 mos. or more	1,069	51.31	5.04	1,069	35.77	4.04	1,070	15.54	2.62
Maternal age									
Less than 20 yrs.	656	48.49	5.65	656	33.08	4.33	656	15.41	2.54
20–29 yrs.	4,134	49.70	5.69	4,134	34.23	4.42	4,134	15.47	2.73
30–39 yrs.	3,512	51.01	5.73	3,512	35.45	4.46	3,522	15.56	2.68
40 yrs. or more	248	51.52	6.25	248	35.89	5.00	247	15.59	2.62
Mother's education									
8th grade or less	382	46.97	5.83	382	31.79	4.56	381	15.18	2.58
9th–12th grades	1,776	48.61	5.68	1,776	33.26	4.36	1,776	15.35	2.70
High school diploma	1,815	49.63	5.62	1,815	34.11	4.47	1,816	15.52	2.66
Voc/technical	187	50.34	5.48	187	34.56	3.86	187	15.78	3.10
Some college	2,099	50.66	5.48	2,099	35.17	4.18	2,098	15.48	2.72
Bachelor's degree	1,371	51.91	5.68	1,371	36.31	4.36	1,370	15.60	2.69
Graduate school (no degree)	157	51.84	5.44	157	36.34	4.09	157	15.50	2.60
Master's degree	551	52.82	5.28	551	36.78	3.99	551	16.05	2.63
Doctoral/professional degree	215	51.99	5.23	215	36.60	3.92	215	15.39	2.64

NOTE: Results were obtained by applying the sampling child weight, W1C0.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

This page is intentionally left blank.

7. PHYSICAL MEASUREMENTS

The following section presents a brief overview of the importance of children's physical growth and development and describes the physical measurements that were obtained. A summary of training and quality control procedures is then presented, as well as correlational evidence of the reliability of the measurements obtained during the 9-month data collection.

Physical 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.

7.1 Overview of the Physical Measurements

In the 9-month data collection, length, weight, and middle upper arm circumference were obtained of all children. Additionally, head circumference was obtained for those children born at very low birth weight, defined as 1,500 grams or less. These measurements were obtained because they are generally recognized as being accurate indicators of children's nutrition, health status, and 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 sponsored by the Center for Disease Control's National Center for Health Statistics. In keeping with this protocol, all physical measurements were obtained twice and composite variables were created to summarize each physical measurement. For more information about how these composites were created, please refer to *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005–013).

7.2 Interviewer Training

During the national training, interviewers had an opportunity to practice obtaining the physical measurements and to demonstrate competence to the trainers. Anyone having questions or

problems about how to obtain the physical measurements was encouraged, and sometimes required, to attend supplementary help labs held in the evenings.

Certification on the physical measurements was obtained during the live practice session described earlier. At the end of the live practice session, interviewers handed in the Child Activity Booklet they used for their practice child. The physical measurements entries were reviewed. Each physical measurement was obtained twice (i.e., child weight 1 and child weight 2 were obtained). If the two measurements obtained were within 5 percent of each other, then the interviewer was certified. If the two measurements were greater than 5 percent different, then the interviewer was required to attend the help lab and also to demonstrate competence to the trainer in obtaining the physical measurements. The purpose of the certification, therefore, was to identify those having problems and make sure they were retrained before leaving training, not to fail those having problems. In this way, by the end of training, all interviewers were certified on the physical measurements.

Interviewers were also instructed during training to evaluate each set of physical measurements to make sure that each set was within the 5 percent difference limit. If the difference was greater than 5 percent, they were instructed to take the physical measurement a third time and record it on the record sheet. At entry, however, if three measurements were recorded, only the two measurements within 5 percent were entered.

On receipt of the first wave of completed cases, it was found that several interviewers made the same mistakes when obtaining the physical measurements. These mistakes included misalignment of the decimal point on the recording sheet; recording child length in inches rather than centimeters; failing to tare the scale used to obtain the mother's and child's weight (resulting in a child weight greater than the mother's weight); failing to obtain a third measure if the first two were more than 5 percent discrepant; or using the wrong side of the head circumference tape. Once these errors were identified, a field memo was sent to all interviewers to notify them about these problems. In addition, those interviewers who were identified as making these errors consistently received a personal update from their supervisors. Once these errors were identified, corrections were made, to the extent that it was possible to do so, prior to data entry. For some cases, however, these errors (for example, failing to obtain a third measure if the first two were more than 5 percent discrepant) could not be corrected retroactively. The analyst is referred to the *User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005-013) for further detail about these corrections.

Table 19 presents the means and standard deviations for these physical measurements, as well as the correlations between the first and second measurements within each set. These statistics are based on the data that were entered before the composites were created. Therefore, these data include those cases in which there may have been a difference greater than 5 percent. In addition, reserve codes[1] have been deleted from these analyses. As a result, the correlation between the first measurement and second measurement probably underestimates the agreement. For further information about how the physical measurements composites were created and how differences larger than 5 percent were treated, please refer to the *User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005-013).

Table 19. Reliability of sets of physical measurements, 9-month data collection: 2001–02

Variable	Mean	Standard deviation	Correlation (r)
Child weight			
C1CHDWT1	9.18 kg	3.13 kg	
C1CHDWT2	8.99 kg	3.62 kg	.81
Child length			
C1CHLNG1	72.68 cm	7.0 cm	
C1CHLNG2	72.42 cm	8.52 cm	.81
Middle upper arm circumference (MUAC)			
MUAC1	15.22 cm	4.40 cm	
C1MUAC2	14.98 cm	4.99 cm	.85
Head circumference			
C1CHHC1	42.96 cm	7.62 cm	
C1CHHC2	42.66 cm	8.61 cm	.84

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

[1] Reserve codes indicate when a value for a variable was not obtained. The reserve code for the physical measurements include −9.

7.3. Physical Measurements in the Nine-Month Data Collection

Details about how the physical measurement components were obtained can be found in the *User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005–013). That document also presents details about relevant contextual variables and how to determine if adjustments to a physical measurement were made, e.g., if only one child weight was obtained.

Because the composite physical measurements are most likely to be used by analysts, they are summarized in table 20.

Table 20. Physical measurement composites variable names, composite description, and measurement unit for the 9-month national data collection: 2001–02

Composite	Composite label	Measurement unit
X1CHWGT	Child's weight	kilograms
X1CHLNG	Child's length	centimeters
X1CHMUAC	Child's middle upper arm circumference	centimeters
X1CHHC	Child's head circumference	centimeters

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

Table 21 summarizes children's average length, weight, and middle upper arm circumference for the key demographic groups, as well as average head circumference when obtained for very low birth weight babies, i.e., birth weight less than or equal to 1,500 grams. To obtain these statistics, reserve codes were deleted and the sampling child weight W1C0 was used.

Table 21. Average physical measurements by key demographic variables, 9-month data collection: 2001–02

Characteristic	Average child physical measurements											
	Weight (kg) (X1CHWGT)			Length (cm) (X1CHLNG)			Middle upper arm circumference (cm) X1CHMUAC			Head circumference (cm) X1CHHC[1]		
	n	Mean	SD	n	Mean	SD	n	Mean	SD	n	Mean	SD
Race/ethnicity												
White	4,180	9.50	1.64	4,243	73.02	4.00	4,159	15.88	1.91	428	44.51	2.51
Black	1,576	9.54	1.60	1,633	73.13	4.16	1,595	16.02	1.99	281	43.92	2.52
Hispanic, race specified	1,414	9.84	.90	1,430	73.28	3.88	1,391	15.83	1.99	143	44.11	2.78
Hispanic, no race specified	634	9.61	1.40	640	73.18	4.02	622	15.80	2.19	62	42.95	3.67
Asian, non-Hispanic	1,091	9.18	1.39	1,102	72.78	3.78	1,076	15.56	1.96	18	44.64	1.50
Native Hawaiian, Pacific Islander	46	9.71	1.44	46	2.90	3.94	45	16.06	1.68	5	43.04	1.24
American Indian, Alaska Native	275	10.19	1.42	275	74.52	4.38	271	16.10	1.90	2	47.35	4.02
More than 1 race, Non-Hispanic	737	9.60	1.82	747	72.91	3.65	736	15.69	1.78	50	44.36	2.34
Poverty status												
Below poverty threshold	2,435	9.53	1.70	2,476	72.99	4.05	2,414	15.86	1.95	283	43.88	2.80
At or above poverty threshold	7,559	9.59	1.66	7,677	73.13	3.97	7,519	15.87	1.97	708	44.27	2.64
Sex												
Male	5,113	9.86	1.56	5,186	73.88	3.88	5,077	16.10	1.95	502	44.55	2.86
Female	4,880	9.28	1.72	4,967	72.28	3.94	4,857	15.63	1.94	489	43.76	2.44
Birth weight												
Normal (2,500 grams or more)	7,359	9.64	1.66	7,457	73.29	3.91	7,291	15.90	1.96	†	†	†
Moderately low (≥ 1,500 and < 2,500 grams	1,540	8.83	1.49	1,572	71.09	3.98	1,540	15.48	1.90	†	†	†
Very low (less than 1,500 grams)	1,057	8.16	1.68	1,087	68.92	4.38	1,066	15.35	1.94	991	44.16	2.69

See notes at end of table.

Table 21. Average physical measurements by key demographic variables, 9-month data collection: 2001–02—Continued

Characteristic	Average child physical measurements											
	Weight (kg) (X1CHWGT)			Length (cm) (X1CHLNG)			Middle upper arm circumference (cm) X1CHMUAC			Head circumference (cm) X1CHHC[1]		
	n	Mean	SD	n	Mean	SD	n	Mean	SD	n	Mean	SD
Age at assessment												
8 mos. or less	351	8.85	1.39	365	70.15	3.10	350	15.58	1.92	15	43.04	3.62
9–10 mos.	5,781	9.28	1.52	5,870	71.86	3.22	5,733	15.86	1.92	321	43.23	2.81
11–12 mos.	2,577	9.81	1.73	2,611	74.02	.43	2,563	15.86	2.02	464	44.44	2.47
13 mos. or more	1,285	10.67	1.71	1,308	77.83	4.01	1,287	16.05	2.02	191	44.92	2.53
Mother's age												
19 years and less	750	9.63	1.76	765	73.05	3.92	739	15.75	1.97	94	44.09	2.33
20–29 years	4,822	9.57	1.65	4,898	73.15	3.99	4,790	15.91	1.91	467	44.01	2.77
30–40 years	4,069	9.58	1.69	4,125	73.08	4.01	4,047	15.86	2.01	388	44.35	2.69
40 years or more	285	9.34	1.44	294	72.51	3.51	287	15.52	2.12	31	43.82	2.63
Mother's education												
8th grade or less	513	9.72	1.73	518	73.43	3.97	501	15.73	1.97	40	3.90	2.96
9th–12th grade	2,115	9.60	1.78	2,145	72.89	4.09	2,079	15.75	2.03	254	44.08	2.86
High school diploma	2,107	9.58	1.69	2,158	73.25	3.98	2,120	15.96	1.98	227	44.21	2.59
Voc/technical	209	9.48	1.71	212	72.62	3.51	211	16.04	1.73	18	44.50	2.15
Some college	2,362	9.57	1.58	2,392	73.32	4.12	2,348	15.92	1.94	239	44.00	2.73
Bachelor's degree	1,576	9.48	1.50	1,601	72.82	3.72	1,566	15.75	1.98	132	44.39	2.69
Grad. school (no degree)	174	9.51	1.51	175	2.43	3.47	171	16.17	1.75	21	44.71	1.32
Master's degree	637	9.64	1.88	645	73.14	3.87	634	16.10	1.82	37	44.35	2.43
Doctoral/professional degree	239	9.70	1.65	242	72.96	4.20	239	15.84	1.95	12	44.78	2.23

† Not applicable.

[1] Obtained of very low birth weight babies only (1,500 grams and less).

NOTE: Results were obtained by applying the sampling child weight W1C0.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

8. THE CHILD OBSERVATIONS

There were two parts to the interviewer-completed observations of child behavior, hereafter referred to as Child Observations: (1) a set of questions about the child's behavior during the administration of the Bayley Short Form–Research Edition (BSF-R); and (2) a set of questions about the child's home environment. The first part was completed by the interviewer in the Child Observations portion of the computer-assisted personal interviewing (CAPI) after the home visit. The second part, assessing key characteristics of 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.

This section describes separately these two parts of the Child Observations and how the items were selected for the ECLS-B. Training procedures to ensure that interviewers understood the observation items are described. Finally, the association between the items describing children's behavior during the BSF-R and their BSF-R outcome scores are presented, as well as descriptive summary statistics of children's home environments as a whole and grouped by key demographic variables.

8.1 Child Observations During the Bayley Short Form–Research Edition

The first part of the Child Observations consisted of a subset of 9 questions from the Behavior Rating Scale (BRS) of the Bayley Scales of Infant Development, Second Edition (BSID-II), 7 of which were completed by the interviewer and 2 of which were completed by the parent. The full BRS consists of 30 items that assess children's behavior during the BSID-II. It is typically used as a clinical tool to interpret children's performance on the BSID-II. For example, a low score on the mental scale could be due to the child's persistent lack of attention to the tasks.

According to the BSID-II manual (Bayley 1993), the BRS items fall into four factors: the attention/arousal factor, the motor quality factor, the orientation/engagement factor, and the emotional regulation factor. 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.

Items for the ECLS-B were selected to sample discrete behaviors that were representative of these four factors, that were developmentally appropriate for the target age, and that were easily rated by field interviewers. That is, items were selected from a range of behaviors that interviewers were likely to see in children of this age in the context of the BSF-R. In addition, items were avoided that were too clinical and therefore too difficult or too subjective for interviewers to observe reliably.

Because only a subset of discrete behaviors from a range of domains are assessed, the BRS items in the ECLS-B should not be considered the same as the BRS. To compare the BRS items in the ECLS-B with the full BRS used in the BSID-II, the analyst is referred to the BSID-II manual for further information. In addition, these items were not selected with the intention of creating a subscale of BRS items.

The BRS items in the ECLS-B at 9 months also included two questions that were asked of the parent respondent at the completion of the BSF-R. These two questions were included in the Child Activity Booklet at the end of the BSF-R section at which point 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.

The first question asked the respondent whether the child's behavior during the BSF-R was typical in terms of whether the child was as alert and active as usual, or 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 analyst is cautioned that 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 analyst can be confident that the child's testing behaviors were representative of the child's general level of functioning.

The second question asked whether the parent respondent thinks the child did as well as he or she could have 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." The interviewer then rated the response in the Child Activity Booklet. 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 analyst is cautioned that 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 (that is, a rating of 3, 4, or 5), then the analyst could have greater confidence that the child's BSF-R scores are representative of the child's general level of functioning.

The interviewer completed the observational items after the home visit was completed and on the basis of child behaviors seen during the BSF-R. After leaving the home, the interviewer accessed these items on the laptop computer. For each item, the interviewer rated the child's behavior during the BSF-R with regard to positive affect (e.g., smiling and laughing); negative affect (e.g., crying and fussing); the amount of interest the child had in the materials; ability to relinquish materials used on one item and accept the material for the next item; ability to focus on the tasks presented; social engagement with the caregiver or interviewer; and the quality of the child's motor control. All items were scored on a well-anchored 5-point scale that incorporates both intensity and frequency of the target behavior. For example, for the item "Positive Affect," the possible ratings are the following:

1. No positive affect displayed;
2. One or two brief displays of positive affect;
3. Three or more brief displays of positive affect;
4. One or two intense, heightened, or prolonged displays of positive affect; or
5. Three or more intense, heightened, or prolonged displays of positive affect.

For more detailed information about the Child Observations items, please see *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005-013).

8.2 Training on the Child Observations

Because interpretation of the observed behaviors could be subjective, these 7 items received a reasonable amount of time and effort during training with extensive use of videotapes and a worksheet

to monitor trainees' understanding of the items. During a 2-hour session, trainees viewed videotapes of the target behaviors sampled from both ends of the rating scale used to evaluate each item. For example, for the item "Child displays positive affect," the trainees saw a videoclip of a child broadly smiling and laughing, and a clip of a child who gives a fleeting and weak, but noticeable, smile. After discussion and completion of the samples on the videotape, trainees completed a worksheet in which they rated behavior samples of the target behaviors on a second videotape. These worksheets were collected and reviewed by the training staff 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.

8.3 Associations of Child Observations with BSF-R Scores

The 7 items included in the Child Observations targeted specific behaviors from the four factors identified in the manual. It was not intended that these items form a scale, but rather offer discrete information helpful for interpreting BSF-R scores. The BSID-II manual suggests three different ways in which the BRS items can be used clinically to interpret children's scores.

One possibility is to sum the scores and use the total score to gain a global impression of the child's behavior. That said, it should be noted that Cronbach's alpha (standardized) for the set of 7 items is .79, which suggests that in the ECLS-B these items, although abbreviated, have conceptual coherence and may be scaled. In addition, the items are modestly to strongly intercorrelated. The correlations between the behavioral items (e.g., positive affect, negative affect, attention to tasks, etc.) range from r = .21 to .61, all significant at p. <.0001. The single motor item, control of movement, also correlates with the behavioral items, ranging from r =.15 to .37, again all significant at p. <.0001. An alternative presented in the manual is to use factor scores, based on the factors presented in the manual. However, too few items from any one factor are included in the ECLS-B for this to be feasible.

Another alternative discussed in the manual is to identify, a priori, particular behaviors that may contribute to children's scores. The manual recommends that a score of 1 indicates a potentially serious impairment in an area, for example, motor quality. For further information, the analyst is referred to chapter 7 of the BSID-II manual.

To demonstrate the association of the Child Observations items with children's scores on the BSF-R, table 22 presents the correlation of each of the observation items with the 9-month BSF-R mental and motor scores. These include the mental scale T-score, motor scale T-score (X1MTLTSC and X1MTRTSC, respectively), which are age-normed and on a 50/10 metric. The mental and motor scale scores (X1MTLSCL and X1MTRSCL, respectively) are true scores that represent the true score the child would have received had the full BSID-II been administered.

Table 22. Correlations of major BSF-R mental and motor scores with the Child Observations items and the two questions for caregivers (n = 10,221), 9-month data collection: 2001–02

	Bayley Short Form–Research Edition Score			
Child Observations	Mental T-score (X1MTLTSC)	Mental scale score (X1MTLSCL)	Motor T-score (X1MTRTSC)	Motor scale score (X1MTRSCL)
Child displays positive affect(R1POSAFF)	.28*	.28*	.18*	.18*
Child displays negative affect (R1NEGAFF)	.19*	.12*	.11*	.05*
Child adapts to change in materials (R1ADAPT)	.21*	.20*	.11*	.11*
Child shows interest in materials (RINTRST)	.32*	.32*	.19*	.20*
Child attends to task (R1ATNTSK)	.32*	.33*	.18*	.19*
Child displays social engagement(R1SOCIAL)	.27*	.26*	.17*	.17*
Child shows control of movements (R1CNTLMV)	.18*	.42*	.29*	.42*
Questions to caregiver				
Caregiver rates child behavior (C1BBHAV1)	.07*	.09*	.08*	.09*
Caregiver rates child's performance (C1PRFRM1)	.11*	.12*	.11*	.11*

* $p < .05$.
NOTE: Results were obtained by applying the sampling child weight, W1C0.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) Nine-Month Data Collection, 2001–02.

8.4 Observations of the Child's Home Environment

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, 2001 and 1979) and from the National Household Education Survey (NHES), also sponsored by 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. The HOME is often used in both academic and large scale surveys to measure key aspects of children's environments, including the quality of parental interaction, the literacy environment, and home environment. In fact, the items from NHES are similar to items from the HOME Short Form, having only some changes in wording and response categories.

The first part of this instrument consists of 8 items about characteristics of the child's environment that are completed in the Child Observation section of CAPI by the interviewer after the home visit is finished. The interviewer records whether or not specific environmental characteristics were observed during the home visit. These items are listed in table 23, along with their means and standard deviations.

These interviewer observation items were complemented by the second part, which consists of 4 questions that were asked of the parent respondent during the parent interview. Although these questions are similar to items in the HOME, they were adopted from NHES with the intention of providing comparability of the ECLS-B with NHES. Although they are similar in content to items in the Short Form of the HOME, their wording and response categories are most consistent with NHES. These items can be found at the bottom of table 23.

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 Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005-013).

Table 23. Summary statistics for home environment set of items (n=10,315), 9-month data collection: 2001–02

Variable	Variable label	Mean	Standard deviation
Child Observations items			
R1RSPKCH	R1 CO165 Caregiver spoke spontaneously to child.	1.07	0.25
R1IORSVB	R1 CO170 Caregiver responded verbally to child.	1.14	0.35
R1IOCRSS	R1 CO175 Caregiver caressed/kissed/hugged child.	1.04	0.21
R1IORSHT[1]	R1 CO180 Caregiver slapped/spanked child.	1.01	0.08
R1IOINTF[1]	R1 CO185 Caregiver interfered with child's actions.	1.22	0.41
R1IOPTYS	R1 CO190 Caregiver provided toys to child.	1.17	0.37
R1IOINVW	R1 CO195 Caregiver kept child in view.	1.02	0.15
R1IOENVS	R1 CO200 Play environment was safe.	1.04	0.19
Parent interview items			
P1READBO[1]	P1 HE102A How often you read to child.	2.23	1.04
P1TELLST[1]	P1 HE102B How often you tell child stories.	2.50	1.11
P1SINGSO[1]	P1 HE102C How often you all sing songs.	1.40	0.74
P1ERRAND[1]	P1 HE102D How often you take child on errands	1.52	0.78

[1] These items were reverse coded.
NOTE: Results were obtained by applying the sampling parent weight, W1R0.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

8.5 Training for HOME Observation and Parent Interview Items

Because the subset of HOME observation items have been used previously in other 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, it was known that it was feasible for interviewers to complete them and that their reliability was

satisfactory. In addition, the HOME items are quite straightforward and presented in their entirety on the CAPI screen. Therefore, training emphasized understanding the items as presented on the CAPI screen, and interviewers were not quizzed on these items or required to pass a certification requirement. Interviewers received training on the remaining four items during the parent interview training and role plays. Further information about training for the parent interview can be found in the *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Design and Operations Report for the Nine-Month Data Collection* (NCES, unpublished report 2004).

8.6 HOME Results from the Nine-Month National Data Collection

The 8 HOME Short Form observation items are coded as "Yes = 1" and "No = 2." This means that 2 negative items must be recoded, R1IORSHT and R1IOINTF, to reverse their direction. The four parent interview items asked parent respondents to report how often in a typical week they (or someone else in the family) engaged in the target activity. Responses were coded as "not at all = 1," "once or twice = 2," "3 to 6 times = 3," and "every day = 4." All four of these items also require recoding to reverse their direction. Means and standard deviations for the HOME-SF observation items and the 4 parent interview items in the ECLS-B 9-month sample are presented in table 23. Reserve codes have been deleted and the parent weight W1R0 was used.

Associations between the HOME items are summarized in table 24, which presents their intercorrelations. Cronbach's alpha (standardized), a measure of internal consistency, for this set of home environment questions is .50, which is low. This suggests that the items do not have a conceptual coherence and are not scalable. However, a principle components factor analysis (using the sampling parent weight, P1C0) with varimax rotation supported 4 factors. Three parent items, P1READBO, P1TELLST and P1SINGSO, loaded most strongly on the first factor, with factor loadings ranging from .62 (P1SINGSO) to .79 (P1READBO) and .80 (P1TELLST). The fourth parent item, P1ERRAND, had only a weak loading on this factor, at .27 (although this was the strongest positive loading for this item). Therefore, the first factor could be characterized as literacy activities. The second factor that emerged could be characterized as the mother's verbal engagement with the child, including, R1RSPKCH, and R1ORSVB, with factor loadings of .84 and .82, respectively. The third factor could be considered parental supervision and consisted of the items R1OENVS, R1OINVW, R1IOCRSS and R1OPTYS,

Table 24. Intercorrelations of home environment items (n = 10,315), 9-month data collection: 2001–02

HOME items											
R1RSPKCH	R1IORSVB	R1IOCRSS	R1IORSHT	R1IOINTF	R1IOPTYS	R1IOINVW	R1IOENVS	P1READBO	P1TELLST	P1SINGSO	P1ERRAND
1.00											
.31*	1.00										
.20*	.17*	1.00									
.06*	.03*	.14*	1.00								
.02*	.02	.04*	.08*	1.00							
.12*	.09*	.20*	.03*	- .03*	1.00						
.15*	.10*	.15*	.11*	.08*	.20*	1.00					
.07*	.05*	.09*	.05*	.07*	.16*	.17*	1.00				
.06*	.06*	.08*	.02	.02*	.13*	.05	.05*	1.00			
.05*	.04*	.05*	.00	.01	.05*	.04*	.00	.50*	1.00		
.05*	.04*	.04*	.01	.02*	.04*	.03*	.04*	.27*	.27*	1.00	
.03*	.02*	.05	-.01	-.01	.06*	.01	-.01	.09*	.09*	.12*	1.00

* P < .05.

NOTE: Results obtained by applying sampling parent weight W1R0.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

with factor loadings of .46, .61, .54, and.70, respectively. Only two items loaded moderately on the fourth factor, R1OINTF and R1ORSHT, with factor loadings of .70 and .60, respectively, with a modest negative load of −.34 for P1ERRAND. Both of these positively loaded items describe parental use of physical interventions to control child behavior. In sum, the low alpha level suggests that these items are not scalable in their entirety. The principal components factor analysis found evidence supporting 4 factors, although some of the factors had as few as two items. The analyst is advised to consider creating an alternative structure, perhaps exploring the scalability of the parents items and the interviewer observations separately.

To examine how the items evaluating the child's home environment performed during the 9-month national data collection, the total score for this set of items was obtained by simply summing the item scores as a scale. The means and standard deviations for the total home environment scale score were obtained for the key demographic variables. These results are presented in table 25.

Table 25. Average total home environment scores by key demographic variables, 9-month data collection: 2001–02

Demographic characteristic	n	Mean	Standard deviation
Total sample	8,613	16.35	2.72
Child race/ethnicity			
White	3,690	15.81	2.55
Black	1,388	17.19	2.84
Hispanic, race specified	1,187	16.90	2.66
Hispanic, no race specified	566	17.22	2.85
Asian, non-Hispanic	920	17.44	3.04
Native Hawaiian/Pacific Islander	36	16.50	2.14
American Indian/Alaska Native	201	16.40	2.53
More than 1 race, non-Hispanic	595	16.04	2.54
Poverty status			
Below poverty threshold	2,002	17.13	2.82
At or above poverty threshold	6,611	16.13	2.66
Sex			
Male	4,369	16.42	2.75
Female	4,244	16.28	2.70
Birth weight			
Normal (2,500 grams or more)	6,343	16.32	2.71
Moderately low (≥ 1,500 and < 2,500 grams	1,341	16.63	2.88
Very low (less than 1,500 grams)	894	16.91	2.85
Child age at assessment			
8 months and less	286	16.25	2.85
9–10 months	4,911	16.39	2.73
11–12 months	2,193	16.47	2.69
13 months or more	1,113	16.02	2.69
Maternal age (years)			
Less than 20	629	16.87	2.58
20–29	4,110	16.53	2.75
30–39	3,559	16.05	2.68
40 or more	253	15.94	2.58

See notes at end of table.

Table 25. Average total home environment scores by key demographic variables, 9-month data collection: 2001–02—Continued

Demographic characteristic	n	Mean	Standard deviation
Mother's education			
8th grade or less	416	18.03	2.93
9–12th grades	1,790	17.05	2.81
High school diploma	1,802	16.67	2.68
Voc/technical	167	15.95	2.56
Some college	2,051	15.97	2.56
Bachelor's degree	1,402	15.65	2.42
Grad. school (no degree)	154	15.25	2.32
Master's degree	568	15.17	2.41
Doctoral/professional degree	207	15.57	2.41
Mother's race/ethnicity			
White	4,028	15.83	2.54
Black	1,400	17.16	2.81
Hispanic, race specified	1,495	17.13	2.76
Hispanic, no race specified	16	15.39	2.36
Asian, non-Hispanic	1,068	17.21	3.02
Native Hawaiian/Pacific Islander	46	16.20	2.48
American Indian/Alaska Native	281	16.36	2.55
More than 1 race, non-Hispanic	215	16.11	32.61

NOTE: Results were obtained by applying the sampling parent weight, W1R0.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

9. INDIRECT ASSESSMENTS OF THE CHILD IN THE PARENT INTERVIEW

The ECLS-B 9-month parent interview also included two sets of questions that obtained indirect assessments of the child's behavior from the parent respondent—a set of questions about when the child reached certain developmental milestones and a set of questions about the child's self-regulation ability and sensorimotor integration. The following sections discuss the rationale for these sets of questions, summarize the items, and present descriptive statistics for key demographic groups.

9.1 Developmental Milestones

The first indirect assessment was a set of four questions that asked the parent respondent about the age at which the child first passed key developmental milestones. Although it is generally accepted that the late achievement of developmental milestones is associated with poorer developmental status and child outcomes in later years, there is little empirical evidence to suggest that the timely achievement or early achievement of developmental milestones has any bearing on future developmental status.

After canvassing the survey and developmental research literatures, it was determined that no appropriate set of questions existed that could be adopted for use in the ECLS-B. Therefore, child development staff reviewed the available published materials to identify key milestones that could be included in the ECLS-B. In particular, it was decided that the milestones should be particularly salient for parents who would be formulating their answers retrospectively. For example, parents would be less likely to remember the age at which a child first used a pincer grasp to pick up a small object than when the child took his first steps. In addition, it was decided that the response options should be straightforward and not lead to embarrassment if the child had not yet passed certain milestones.

After reviewing the contents and response formats of several screening instruments, it was decided that several of the behaviors contained in the *Child Development Inventory* (Ireton 1997) would be useful. In addition, the response options were simplified somewhat to enable respondents to report that the child had not yet achieved a certain milestone. The results of field testing determined that the

following items most successfully obtained information about the age at which the child first performed the milestone:

- P1AGSIT: How old was child in months when he/she started to sit alone, steady, without support?
- P1AGCRWL: How old was child in months when he/she started to crawl on hands and knees?
- P1AGSTND: How old was child in months when he/she started to pull him/herself to a standing position?
- P1AGWALK: How old was child in months when he/she started to first walk while holding onto something, such as furniture?

Table 26 presents the average age at which the sample children, who have achieved that milestone, passed each developmental milestone, grouped by key demographic variables. These are demographic characteristics where differences associated with the emergence of these skills are considered important to inform educational research. Because these averages only include the children who have passed the milestone in question, they will slightly underestimate the ages at which the milestones were reached. These averages do not include children who have yet to achieve the milestone. For further information, please refer to chapter 5 of *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Methodology Report for the Nine-Month Data Collection, Volume 1: Sampling* (NCES 2005-113), although no comparable national norms are available. Only 3.6 percent of children were not yet sitting; 15.1 percent were not yet crawling; 17 percent were not yet pulling to a stand, and 26.3 percent were not yet walking with help. To obtain these statistics, all reserve codes were deleted and the data were weighted using the parent weight W1R0.

Table 26. Average age (and standard deviation) of children who have passed milestones in decimal months, 9-month data collection: 2001–02

	Average age at milestone in decimal months (and standard deviation)											
	Sitting alone (P1AGSIT)			Crawling (P1AGCRWL)			Pull to standing (P1AGSTND)			Walk with help (P1AGWALK)		
Characteristic	n	Mean	SD	n	Mean	SD	n	Mean	SD	n	Mean	SD
Child race/ethnicity												
White	4,259	6.04	1.28	3,642	7.31	1.41	3,519	8.04	1.38	3,078	8.67	1.38
Black	1,624	5.88	1.42	1,511	6.79	1.44	1,492	7.71	1.38	1,349	8.27	1.39
Hispanic, race specified	1,456	6.25	1.46	1,271	7.10	1.59	1,244	8.04	1.57	1,129	8.66	1.48
Hispanic, no race specified	633	6.32	1.35	534	7.18	1.48	544	8.12	1.47	475	8.72	1.41
Native Hawaiian/ Pacific Islander	48	6.51	1.10	46	6.86	1.48	45	7.75	1.30	39	8.38	1.41
Asian, non-Hispanic	1,167	6.44	1.29	1,061	7.16	1.38	1,032	8.23	1.29	904	8.82	1.29
American Indian/ Alaska Native	288	6.14	1.42	274	7.33	1.52	262	8.23	1.63	245	8.92	1.58
More than 1 race, non-Hispanic	755	5.84	1.32	685	6.99	1.46	667	7.73	1.40	614	8.49	1.46
Poverty status												
Below poverty threshold	2,494	6.18	1.47	2,224	6.94	1.50	2,194	7.95	1.50	1,950	8.52	1.49
At or above poverty threshold	7,774	6.06	1.34	6,832	.24	1.45	6,648	8.01	1.40	5,913	8.64	1.39
Sex												
Male	5,209	6.12	1.36	4,617	7.15	1.46	4,499	7.97	1.41	4,015	8.60	1.40
Female	5,059	6.05	1.33	4,439	7.20	1.47	4,343	8.02	1.45	3,848	8.62	1.43
Birth weight												
Normal (2,500 grams or more)	7,740	6.02	1.30	6,960	7.13	1.45	6,895	7.95	1.41	6,210	8.57	1.40
Moderately low (≥ 1,500 and < 2,500 grams)	1,573	6.71	1.45	1,343	7.61	1.50	1,266	8.41	1.47	1,111	8.99	1.46
Very low (less than 1,500 grams)	914	8.21	1.77	717	9.00	1.75	646	9.66	1.70	509	10.22	1.66
Maternal age												
Less than 20 yrs.	750	5.90	1.33	675	6.97	1.52	678	7.79	1.46	619	8.37	1.41
20–29 yrs.	4,928	6.03	1.36	4,440	7.07	1.45	4,349	7.91	1.42	3,926	8.54	1.40
30–39 yrs.	4,216	6.16	1.32	3,618	7.33	1.46	3,515	8.12	1.41	3,063	8.72	1.42
40 or more yrs.	303	6.36	1.38	254	7.50	1.34	241	8.14	1.31	203	8.76	1.16

See notes at end of table.

Table 26. Average age (and standard deviation) of children who have passed milestones in decimal months, 9-month data collection: 2001–02—Continued

Characteristic	Average age at milestone in decimal months (and standard deviation)											
	Sitting alone (P1AGSIT)			Crawling (P1AGCRWL)			Pull to standing (P1AGSTND)			Walk with help (P1AGWALK)		
	n	Mean	SD	n	Mean	SD	n	Mean	SD	n	Mean	SD
Mother's race/ethnicity												
White	4,670	6.03	1.28	3,998	7.30	1.42	3,881	8.03	1.38	3,417	8.65	1.39
Black	1,651	5.89	1.40	1,531	6.78	1.43	1,514	7.69	1.39	1,368	8.27	1.40
Hispanic, race specified	1,779	6.30	1.43	1,536	7.14	1.54	1,517	8.07	1.53	1,354	8.69	1.45
Hispanic, no race specified	20	6.75	1.60	15	7.01	1.17	13	7.65	1.06	12	8.71	1.46
Asian, non-Hispanic	1,354	6.42	1.29	1,224	7.18	1.40	1,189	8.21	1.30	1,042	8.82	1.33
Native Hawaiian/ Pacific Islander	55	6.26	1.43	56	6.95	1.47	52	7.83	1.55	49	8.40	1.52
American Indian/ Alaska Native	390	5.97	1.38	369	7.23	1.50	361	8.08	1.61	337	8.80	1.50
More than 1 race, non-Hispanic	273	5.84	1.27	254	7.02	1.53	251	7.69	1.31	229	8.46	1.26
Mother's education												
8th grade or less	517	6.32	1.41	451	7.11	1.52	443	8.02	1.49	397	8.64	1.47
9–12th grades	2,153	6.15	1.40	1,931	6.99	1.50	1,895	7.93	1.48	1,733	8.57	1.47
High school diploma	2,166	6.07	1.41	1,956	6.99	1.49	1,923	7.92	1.46	1,708	8.57	1.44
Vocational/ technical	220	5.97	1.30	184	7.12	1.35	180	7.86	1.26	160	8.46	1.13
Some college	2,412	6.01	1.31	2,116	7.19	1.44	2,081	7.94	1.40	1,848	8.57	1.39
Bachelor's degree	1,636	6.06	1.29	1,411	7.52	1.38	1,359	8.16	1.33	1,190	8.69	1.32
Graduate school (no degree)	182	6.06	1.21	153	7.20	1.32	146	7.94	1.62	127	8.50	1.37
Master's degree	668	6.05	1.21	572	7.64	1.34	550	8.28	1.23	467	8.86	1.36
Doctoral/ professional degree	249	6.05	1.17	219	7.43	1.28	212	8.01	1.35	187	8.68	1.23
Milestone not yet achieved (%)			3.6			15			17			26

NOTE: These results were obtained by applying the sampling parent weight W1R0.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine Month Data Collection, 2001–02.

9.2 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 measure is a screener that was designed to be completed by parents and obtains information about children's self-regulatory behaviors. It is based on research that has shown that sensorimotor integration problems in young children may be associated with later attentional problems, such as attention deficit hyperactivity disorder (ADHD). For children with sensorimotor integration problems, sensory information, such as touch, sound, and movement, is misinterpreted for various neurophysiological reasons. This 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).

The items selected for the ECLS-B at 9 months were chosen on the basis of their ability to identify children with sensorimotor and self-regulatory difficulties that may lead to problems with attention in the preschool years and later. However, it was also deemed important to select items that were salient to parents and easily comprehensible.

The parent respondent was asked to report whether the child was "never," "used to be," "sometimes," or "most times" like the target item. Items were selected that were easy for parent respondents to understand and to endorse. Items were selected from four domains including self-regulation, attention, sleep, and reactivity to sound. The items include the following:

- P1FUSSY: Child is frequently irritable or fussy;
- P1WHMPR: Child goes easily from a whimper to an intense cry;
- P1ATTN: Child demands your attention and company constantly;
- P1WAKES: Child wakes up 3 or more times in the night and is unable to go back to sleep;
- P1HLPSLP: Child needs a lot of help to fall asleep; (e.g., rocking, long walks, stroking hair, car rides, etc);
- P1STRTL: Child startles or is upset by loud sounds such as a vacuum, doorbell, or barking dog; and
- P1NOWAIT: Child is unable to wait for food or toys without crying or whining.

According to the ITSC manual, although the items are presented in conceptually oriented categories, there is just one factor that corresponds most closely to self-regulation. Hence, the ITSC score is obtained simply by summing the item responses, with "Never = 0," "Used to be = 1," "Sometimes = 2," and "Always = 3." ECLS-B field testing, however, suggested two factors, the first factor being self-regulation and the second factor being sleep problems. In addition, the field test found that one item, "startles easily," loaded weakly on both factors. Coefficient alpha (standardized) for the above ITSC items, using the complete 9-month data set (weighted), was .63, which is adequate. This suggests that, although the items have consistency, the field test finding of two factors is reasonable. The analyst is advised to examine the data carefully before assuming that the ITSC, as implemented in the ECLS-B, is composed of a unitary factor. Table 27 presents the mean ITSC score by groups, based on the key demographic variables considered to be potentially influential on the development of children's self-regulation behavior. To obtain these statistics, reserve codes were deleted and the parent weight W1R0 was used.

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 consider prorating the summed scores and determine a prorated cut-off score. Therefore, the analyst is referred to the ITSC manual (DeGangi, Poisson, Sickel, and Wiener 1995) for further information as well as to *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005-013).

Table 27. Average scores (and standard deviations) for self-regulation behaviors by key demographic characteristics, 9-month data collection: 2001–02

Demographic characteristic	n	Infant/Toddler Symptom Checklist (ITSC) self-regulation behavior															
		Total score		Frequently fussy (P1FUSSY)		Whimper to cry easily (P1WHMPR)		Demands attention (P1ATTN)		Wakes 3+ in the night (P1WAKES)		Needs help to fall asleep (P1HLPSLP)		Startles easily (P1STRTL)		Unable to wait (P1NOWAIT)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Maternal race/ ethnicity																	
White	4,860	7.87	3.79	1.19	0.98	1.07	1.07	1.82	1.09	0.45	0.79	0.94	1.14	0.81	1.05	1.59	1.02
Black	1,723	8.81	3.85	1.51	0.94	1.20	1.04	2.16	1.04	0.48	0.80	0.89	1.10	0.97	1.14	1.59	1.07
Hispanic, race specified	1,869	7.99	3.82	1.16	1.04	1.04	1.07	2.10	1.05	0.48	0.85	0.87	1.13	0.86	1.11	1.49	1.11
Hispanic, no race specified	21	7.89	4.88	1.08	1.01	1.07	1.16	2.15	1.02	0.38	0.72	0.99	1.21	0.94	1.02	1.28	1.14
Asian, non-Hispanic	1,384	8.72	3.95	1.24	0.98	1.13	1.04	2.05	1.05	0.70	0.99	1.09	1.19	1.06	1.16	1.46	1.10
Native Hawaiian/ Pacific Islander	57	.31	3.40	1.40	0.84	1.66	0.85	2.01	1.07	0.63	0.87	0.80	1.05	1.07	1.13	1.70	0.99
American Indian/ Alaska Native	398	8.41	3.67	1.50	0.92	1.08	1.04	2.14	1.02	0.50	0.80	0.89	1.11	0.77	1.04	1.53	1.02
More than 1 race, non-Hispanic	280	8.04	4.18	1.29	1.03	1.25	1.09	1.95	1.81	0.51	0.79	0.84	1.15	0.72	1.01	1.48	1.10
Poverty status																	
Below poverty threshold	2,597	8.45	4.00	1.34	1.01	1.18	1.09	2.15	1.06	0.48	0.84	0.85	1.12	0.88	1.12	1.57	1.10
At or above poverty threshold	8,075	7.96	3.78	1.20	1.00	1.06	1.06	1.89	1.08	0.47	0.81	0.94	1.14	0.84	1.07	1.56	1.04

See notes at end of table.

Table 27. Average scores (and standard deviations) for self-regulation behaviors by key demographic characteristics, 9-month data collection: 2001–02

Demographic characteristic	n	Infant/Toddler Symptom Checklist (ITSC) self-regulation behavior															
		Total score		Frequently fussy (P1FUSSY)		Whimper to cry easily (P1WHMPR)		Demands attention (P1ATTN)		Wakes 3+ in the night (P1WAKES)		Needs help to fall asleep (P1HLPSLP)		Startles easily (P1STRTL)		Unable to wait (P1NOWAIT)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Sex																	
Male	5,451	8.20	3.85	1.26	1.00	1.10	1.06	1.94	1.09	0.52	0.85	0.96	1.15	0.84	1.08	1.60	1.05
Female	5,221	7.93	3.81	1.21	1.00	1.07	1.07	1.96	1.08	0.43	0.78	0.88	1.12	0.86	1.09	1.52	1.06
Child age at assessment																	
8 mos. or less	366	7.83	4.00	1.13	1.03	0.95	1.07	1.94	1.13	0.59	0.96	0.95	1.14	0.74	1.05	1.53	1.07
9–10 mos.	5,896	8.08	3.84	1.20	1.00	1.07	1.06	1.93	1.09	0.49	0.82	0.97	1.15	0.85	1.09	1.56	1.06
11–12 mos.	2,630	8.07	3.75	1.25	1.00	1.13	1.07	1.94	1.06	0.44	0.78	0.89	1.13	0.87	1.08	1.56	1.04
Greater than 12 mos.	1,325	8.39	3.87	1.41	0.98	1.18	1.07	2.06	1.06	0.47	0.83	0.78	1.10	0.86	1.10	1.62	1.04
Birth weight (grams)																	
Normal (2,500 or more)	7,832	8.04	3.82	1.23	1.00	1.07	1.07	1.94	1.08	0.47	0.82	0.92	1.14	0.84	1.08	1.56	1.06
Moderately low (≥ 1,500 and < 2,500	1,645	8.43	4.00	1.30	1.00	1.24	1.08	1.98	1.10	0.48	0.80	0.93	1.14	0.93	1.12	1.57	1.06
Very low (less than 1,500)	1,153	8.32	3.95	1.27	1.00	1.14	1.07	1.97	1.14	0.45	0.81	0.87	1.11	1.04	1.14	1.58	1.07
Maternal age																	
Less than 20 yrs.	774	8.82	3.88	1.37	1.00	1.24	1.08	2.21	0.98	0.47	0.80	1.03	1.15	0.86	1.08	1.65	1.03
20–29 yrs.	5,132	8.25	3.71	1.27	1.01	1.12	1.06	2.07	1.04	0.45	0.79	0.90	1.14	0.83	1.09	1.60	1.05
30–39 yrs.	4,372	7.74	3.92	1.17	0.98	1.02	1.07	1.75	1.12	0.50	0.84	0.94	1.14	0.86	1.08	1.50	1.06
40 yrs. or more	319	7.45	4.20	1.04	1.01	1.09	1.07	1.75	1.14	0.54	0.90	0.75	1.10	0.86	1.08	1.41	1.10

See notes at end of table.

Table 27. Average scores (and standard deviations) for self-regulation behaviors by key demographic characteristics, 9-month data collection: 2001–02

Demographic characteristic	Infant/Toddler Symptom Checklist (ITSC) self-regulation behavior																	
		Total score		Frequently fussy (P1FUSSY)		Whimper to cry easily (P1WHMPR)		Demands attention (P1ATTN)		Wakes 3+ in the night (P1WAKES)		Needs help to fall asleep (P1HLPSLP)		Startles easily (P1STRTL)		Unable to wait (P1NOWAIT)	
	n	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Mother's education																	
8th grade or less	547	7.84	4.25	1.13	1.07	1.02	1.07	2.08	1.05	0.51	0.87	0.72	1.09	0.94	1.16	1.44	1.17
9–12th grade	2,254	8.75	3.85	1.40	0.98	1.20	1.08	2.18	1.04	0.51	0.87	0.97	1.16	0.86	1.11	1.63	1.08
High school. diploma	2,274	8.24	3.80	1.33	0.98	1.15	1.06	2.04	1.08	0.45	0.79	0.88	1.12	0.83	1.08	1.56	1.06
Vocational/ technical	224	8.04	3.83	1.29	0.94	0.95	1.04	2.01	1.04	0.45	0.78	0.97	1.16	0.85	1.06	1.52	1.06
Some college	2,496	7.76	3.72	1.14	1.00	1.02	1.05	1.88	1.08	0.42	0.76	0.89	1.13	0.84	1.07	1.56	1.03
Bachelor's degree	1,684	7.71	3.67	1.15	0.98	1.02	1.07	1.75	1.06	0.48	0.82	0.95	1.14	0.80	1.04	1.55	1.02
Graduate school (no degree)	182	7.26	4.10	1.00	0.96	1.18	1.14	1.48	1.15	0.42	0.75	0.93	1.15	0.89	1.08	1.40	1.07
Master's degree	686	7.78	3.87	1.11	0.97	0.97	1.08	1.61	1.08	0.59	0.86	1.09	1.14	0.93	1.11	1.48	1.00
Doctoral/ Professional degree	256	7.46	3.81	0.84	0.94	1.07	1.02	1.56	1.06	0.63	0.93	1.00	1.14	0.84	1.03	1.51	0.97

NOTE: These results were obtained by applying the sampling parent weight, W1R0.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

This page is intentionally left blank.

10. OTHER INSTRUMENTS IN THE PARENT INTERVIEW

The 9-month parent interview included several sets of questions that assessed various aspects of children's home environments and parents' attitudes toward and knowledge about children. This section introduces each of these measures and presents psychometric data obtained from the 9-month data collection. For further information about the training of interviewers to conduct the parent interview, please refer to *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) Design and Operations Report for the Nine-Month Data Collection* (NCES, unpublished report 2004). These measures include a set of items that obtain information about the child's home environment and everyday activities, a set of items that measure parent respondents' knowledge about age-appropriate child development, and, finally, a set of items that assess parents' authoritative versus authoritarian parenting beliefs. Each of these sets of questions is discussed in turn in the following sections.

10.1 Set of Questions From the Knowledge of Infant Development Inventory

This subset of questions consisted of 11 items selected from the much larger Knowledge of Infant Development Inventory (KIDI)(MacPhee 1981). This instrument was designed to assess knowledge of parents' caregiving practices, their knowledge of developmental processes, and their knowledge of common infant norms of behavior. The author of the KIDI recommended the particular items that were important to include for children of the target age of the 9-month data collection. Four of these items are simple items that present a statement about children's characteristics or abilities that may be correct or incorrect. The respondent was asked to indicate agreement or disagreement with the statement. Another 7 items can be considered compound items in which the respondent first indicated agreement or disagreement with the statement and, in the case of disagreement, indicated whether it would be characteristic of a younger or older child.

The four simple items for which respondents indicated they agreed or disagreed included the following:

- P1AMTSLP: All infants need the same amount of sleep.
- P1SIBWET: A young brother or sister may start wetting the bed or thumbsucking when a new baby arrives in the family.

- P1CHSPCR: A child thinks he is speaking correctly even when he says words and sentences in an unusual or different way, like “I goed to town” or “What the dollie have?”

- P1CHLGCP: Children learn all of their language by copying what they have heard adults say.

The 7 compound questions consist of two parts. For the first part, the respondent indicated agreement or disagreement with the statement. For the second part, if the respondent disagreed, she then indicated whether the statement was typical of a younger child or an older child.

- P1RIGHTWR: A one-year-old knows right from wrong. Do you agree or disagree? If disagree,

  - P1OYRGWR: Would a child be younger or older than one year when she first knows right from wrong?

- P1KNWSNM: A baby will begin to respond to her name at 10 months. Do you agree or disagree? If disagree,

  - P1OYKNNM: Would a child be younger or older than 10 months when she first responds to her name?

- P1TLTTRN: Most infants are ready to be toilet trained by one year of age. Do you agree or disagree? If disagree,

  - P1OYTLTR: Would that be a younger or an older child?

- P1REMTOY: A baby of 12 months can remember toys he has watched being hidden. Do you agree or disagree? If disagree,

  - P1OYRMTY: Would a baby be younger or older than 12 months when he first remembers toys he has watched being hidden?

- P1SHRPLY: One-year-olds often cooperate and share when they play together. Do you agree or disagree? If disagree,

  - P1OYSHPL: Would children be younger or older than one year when they often cooperate and share when they play together?

- P1GBRRCH: A baby is about 7 months old before she can reach for and grab things. Do you agree or disagree? If disagree,

  - P1OYGRRC: Would a baby be younger or older than 7 months before she can reach for and grab things?

- P1WRD1ST: A baby usually says his first real word by six months of age. Do you agree or disagree? If disagree,
    - P1OY1SWD: Would a baby be younger or older than six months when he says his first real word?

Each response is scored as correct or incorrect and the sum of the correct scores was obtained. These scoring rules are summarized in table 28.

For further details about the KIDI items in the public data set and how to score them, the analyst is referred to the *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005-013). A total score for correct responses to the KIDI subset of items was obtained for each case. Table 29 presents the average total KIDI scores by subgroups, based on key demographic variables thought to be influential in respondents' knowledge of child development, including maternal race/ethnicity, maternal age, maternal education, poverty status, child gender, child age at assessment, and child birth weight. This total score was obtained by applying the scoring instructions presented in the *User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book*. In order for the analyst to reproduce these results, it should be remembered that the reserve codes are deleted.

Table 28. Scoring key for KIDI items in the parent interview and resident father questionnaire, 9-month data collection: 2001–02

Variable name	Variable description	Correct response option and points assigned
Parent CAPI instrument		
P1AMTSLP	P1 PA015A Infants need same amount of sleep.	(2) Disagree (1 point)
P1SIBWET	P1 PA015B Young siblings may wet bed.	(1) Agree (1 point)
P1CHSPCR	P1 PA015C Child thinks speaks correctly.	(1) Agree (1 point)
P1CHLGCP	P1 PA015D Children learn all language by copying.	(2) Disagree (1 point)
P1RIGHTWR	P1 PA020A 1 Year old knows right from wrong.	(2) Disagree
P1OYRGWR	P1 PA020B Older/younger learns right from wrong.	(2) Older (1 point)
P1KNWSNM	P1 PA025A Baby responds to name at 10 months.	(2) Disagree
P1OYKNNM	P1 PA025B Older/younger baby knows name.	(1) Younger (1 point)
P1TLTTRN	P1 PA030A Child is ready for toilet training at 1 year	(2) Disagree
P1OYTLTR	P1 PA030B Older/younger ready for toilet training.	(2) Older (1 point)
P1REMTOY	P1 PA035A 1 Year old can remember hidden toys.	(1) Agree (1 point)
P1OYRMTY	P1 PA035B Older/younger remembers hidden toys.	
P1SHRPLY	P1 PA040A 1 Year-olds share/play together.	(2) Disagree
P1OYSHPL	P1 PA040B Older/younger share/play together.	(2) Older (1 point)
P1GBRRCH	P1 PA045A Baby can grab/reach at 7 months.	(2) Disagree
P1OYGRRC	P1 PA045B Older/younger can grab/reach.	(1) Younger (1 point)
P1WRD1ST	P1 PA050A Baby says first word by 6 months.	(2) Disagree
P1OY1SWD	P1 PA050B Older/younger says first word.	(2) Older (1 point)
Resident father questionnaire		
F1RGHTWR	F1 Q7A 1 Year-old knows right from wrong.	(2) Older (1 point)
F1KNWSNM	F1 Q7B Baby responds to own name at 10 months.	(3) Younger (1 point)
F1TLTTRN	F1Q7C Child is ready for toilet training at 1 year	(2) Older (1 point)
F1REMTOY	F1 Q7D 1 Year-old can remember hidden toys.	(1) Agree (1 point)
F1SHRPLY	F1 Q7E 1 Year-olds share/play together.	(2) Older (1 point)
F1GRBRCH	F1 Q7F Baby can grab/reach at 7 months.	(3) Younger (1 point)
F1WRD1ST	F1 Q7G Baby says first word by 6 months.	(2) Older (1 point)

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort, Nine-Month Data Collection, 2001–02.

Table 29. Average KIDI subset total scores by key demographic characteristics, 9-month data collection: 2001–02

Demographic characteristic	n	Average total score	Standard deviation
Maternal race/ethnicity			
White	4,870	7.57	1.87
Black	1,724	5.78	1.95
Hispanic, race specified	1,870	5.29	2.16
Hispanic, no race specified	21	3.62	1.36
Asian, non-Hispanic	1,388	5.71	2.13
Native Hawaiian/Pacific Islander	57	6.43	2.05
American Indian/Alaska Native	398	6.65	1.99
More than 1 race, non-Hispanic	280	7.19	1.85
Poverty status			
Below poverty threshold	2,603	5.65	2.13
At/above poverty threshold	8,085	7.04	2.13
Sex			
Male	5,460	6.71	2.18
Female	5,228	6.73	2.24
Birth weight			
Normal (2,500 grams or more)	7,844	6.73	2.21
Moderately low ($\geq$ 1,500 and < 2,500 grams	1,647	6.50	2.17
Very low (less than 1,500 grams)	1,155	6.50	2.09
Child's age at assessment			
8 months or less	367	6.47	2.41
9–10 months	5,904	6.84	2.24
11–12 months	2,634	6.65	2.15
13 months or more	1,325	6.52	2.02
Maternal age			
Less than 20 yrs.	774	5.66	1.86
20–29 yrs.	5,708	6.52	2.16
30–39 yrs.	3,812	7.22	2.22
40 yrs. or more	319	7.23	2.35

See notes at end of table.

Table 29. Average KIDI subset total scores by key demographic characteristics, 9-month data collection: 2001–02—Continued

Demographic characteristic	n	Average total score	Standard deviation
Mother's education			
8th grade or less	549	4.53	2.00
9th–12th grades	2,258	5.76	2.02
High school diploma	2,227	6.40	2.05
Voc/technical	224	6.81	1.99
Some college	2,497	7.23	1.96
Bachelor's degree	1,688	7.80	1.96
Graduate school (no degree)	182	8.14	2.13
Master's degree	687	8.00	1.89
Doctoral/professional degree	257	7.81	2.09

NOTE: Weighted data, W1R0 used.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

10.2 Parenting Attitudes Questions

Associations have been found between parenting attitudes and children's early developmental outcomes, particularly with respect to authoritarian versus authoritative parenting attitudes (Baumrind 1966; and for a review, please see Maccoby and Martin 1983). The contrast between authoritarian and authoritative parenting has been demonstrated to be a robust finding in the child development literature, and the effects of these styles have been found to be enduring, with significant associations with key child outcome measures through adolescence and beyond (e.g., Baumrind 1991).

In order to capture information about the parent respondent's parenting attitudes, 5 items were included that demonstrated common beliefs or practices that are associated with authoritative and authoritarian parenting. Each item consists of a pair of statements and the parent indicated which of the two she agreed with was closest to her own ideas. The instructions to respondents and items included the following::

Please choose the one that is closest to your own ideas, either statement 1 or 2.

- **P1PCKCRY:**

  1. You can spoil a tiny baby by picking him up every time he cries, or

2. You cannot spoil a tiny baby by picking him up every time he cries.

- **P1TOITRN:**

  1. If a mother trains her baby properly, he will not need diapers after he is one year old, or

  2. It is better not to start toilet training a baby until he is at least a year old.

- **P1FEEDSC:**

  1. Small babies should be fed when they are hungry, or

  2. Small babies should be fed on a regular schedule.

- **P1HBVSHP:**

  1. It is important to see that a young child does not form bad habits, or

  2. If a young child is happy, he will not form bad habits.

- **P1MOMVAL:**

  1. Most mothers nowadays let their children get away with too much, or
  2. Most mothers nowadays do a pretty good job of raising their children.

Analysts interested in knowing more about how these items were administered in the parent interview are referred to the *ECLS-B Design and Operations Report for the Nine-Month Data Collection.* A total score was obtained by summing the responses to these items. Each of these items is unidirectional in that a higher score indicates authoritarian beliefs or practices. Results from the fall 2000 field test showed that these items successfully discriminate authoritarian and authoritative parenting beliefs.

Table 30 presents group differences obtained by summing the scores of the parenting beliefs questions. Reserve codes have been deleted and the data were weighted using the parent weight, W1R0. Grouping variables are the same as above and are believed to be most readily associated with parenting beliefs.

Table 30. Average total scores on authoritarian/authoritative parenting beliefs by key demographic characteristics, 9-month data collection: 2001–02

Demographic characteristic	n	Average total score	Standard deviation
Maternal race/ethnicity			
White	4,706	7.57	0.95
Black	1,690	7.45	1.08
Hispanic, race specified	1,819	7.63	1.12
Hispanic, no race specified	19	7.71	0.76
Asian, Non-Hispanic	1,314	7.82	1.07
Native Hawaiian/Pacific Islander	57	7.45	1.21
American Indian/Alaska Native	386	7.59	1.05
More than 1 race, non- Hispanic	274	7.58	1.00
Poverty status			
Below poverty threshold	2,530	7.46	1.09
At/above poverty threshold	7,814	7.61	0.99
Sex			
Male	5,285	7.57	1.02
Female	5,064	7.57	1.01
Birth weight			
Normal (2,500 grams or more)	7,597	7.57	1.01
Moderately low ($\geq$ 1,500 and < 2,500 grams	1,587	7.54	1.03
Very low (less than 1,500 grams)	1,118	7.62	1.06
Child age at assessment			
8 months or less	356	7.57	1.03
9–10 months	5,710	7.56	1.02
11–12 months	2,556	7.59	1.01
13 months or more	1,291	7.59	0.99
Maternal age			
Less than 20 yrs.	756	7.35	1.09
20-29 yrs.	4,993	7.47	1.02
30-39 yrs.	4,219	7.74	0.97
40 yrs. or more	302	7.67	0.95

See notes at end of table.

Table 30 Average total scores on authoritarian/authoritative parenting beliefs by key demographic characteristics, 9-month data collection: 2001–02—Continued

Demographic characteristic	n	Average total score	Standard deviation
Mother's education			
8th grade or less	529	7.60	1.11
9–12th grades	2,197	7.43	1.10
High school diploma	2,215	7.44	1.01
Voc/technical	219	7.51	0.96
Some college	2,423	7.60	0.97
Bachelor's degree	1,625	7.78	0.94
Graduate school (no degree)	174	7.83	0.88
Master's degree	651	7.84	0.89
Doctoral/professional degree	243	7.91	0.86

NOTE: Weighted data, W1R0 used.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

This page is intentionally left blank.

REFERENCES

Aksan, N. and Kochanska, G. (2004). Links between systems of inhibition from infancy to preschool years. *Child Development, 75*: 1477–1499.

Ayres, A.J. (1979). *Sensory Integration and the Child.* Los Angeles: Western Psychological Services.

Barnard, K.E. (1997). Influencing parent-child interactions for children at risk. In M. J. Guralnick (Ed.), *The Effectiveness of Early Intervention*. Baltimore, MD: Paul Brooks.

Baumrind, D. (1966). Effects of authoritative parental control on child behavior. *Child Development, 37*(4): 887–907.

Baumrind, D. (1991). The influence of parenting style on adolescent competence and substance use. *Journal of Early Adolescence, 11*(1): 56–95.

Bayley, N. (1993). *Bayley Scales of Infant Development, Second Edition*. San Antonio, TX: The Psychological Corporation.

Bethel, J., Green, J.L., Kalton, G., and Nord, C. (Forthcoming). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Methodology Report for the Nine-Month Data Collection (2001–02), Volume 2: Sampling* (NCES 2005–113). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Bornstein, M.H. and Seuss, P.E. (2000). Physiological self-regulation and information processing in infancy: Cardiac vagal tone and habituation. *Child Development, 71*: 273–287.

Bornstein, M.H., and Sigman, M.D. (1986). Continuity in mental development from infancy. *Child Development, 57*: 251–274.

Burns, W.J., Burns, K.A., and Kabacoff, R.I. (1992). Item and factor analyses of the Bayley Scales of Infant Development. In C. Rovee-Collier & L.P. Lipsitt (Eds.), *Advances in Infancy Research, 7*, 199–214.

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, Third Edition, 2001*. Little Rock, AR: University of Arkansas..

DeGangi, G.A., Poisson, S., Sickel, R.Z., and Wiener, A.S. (1995). *Infant/Toddler Symptom Checklist.* San Antonio, TX: The Psychological Corporation.

Embretson, S.E., and Reise, S.P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Erlbaum Publishers.

Fisher, A.G., Murray, E.A., and Bundy, A.C. (1991). *Sensory Integration Theory and Practice.* Philadelphia: F.A. Davis.

Flynn, J. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*: 29–51.

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* (Working Paper No. 97-24). U.S. Department of Education, Washington, DC: National Center for Education Statistics.

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*. Newbury Park. CA: Sage Publications, 109–110.

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.

Maccoby, E.E., and Martin, J.A. (1983). Socialization in the context of the family: Parent–child interaction. In P. H. Mussen (Ed.) & E. M. Hetherington (Vol. Ed.), *Handbook of child psychology: Vol. 4. Socialization, personality, and social development* (4th ed., pp. 1–101). New York: Wiley.

MacPhee, D. (1981). *Knowledge of Infant Development Inventory Manual.* Unpublished manual, University of Denver, Denver, CO.

Meisels, S.J., Atkins-Burnett, S., and Nicholson, J. (1996). *Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children* (Working Paper No. 96-18). U.S. Department of Education, Washington, DC: National Center for Education Statistics.

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 Menestral, S., Zaslow, M., Dungee Greene, A., Mariner, C., Romano, A., and Bridges, L. (1999). *A Birth Cohort Study: Conceptual and Design Considerations and Rationale* (Working Paper No. 1999-01). 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. (2005). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), User's Manual for the ECLS-B Nine-Month Public-Use Data File and Electronic Code Book* (NCES 2005–013). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Nunnally, J.C. (1978) *Psychometric Theory*. P. 245. New York: McGraw-Hill Book Company.

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.

Raver, C.C. (2004). Placing emotional self-regulation in sociocultural and socioeconomic contexts. *Child Development, 75*: 346–353.

Rock, D., and Pollack, J. (2002). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Shonkoff, J.P., Hansen-Cram, P., Krauss, M.W., and Upshur, C.C. (1992). *Development of infants with* disabilities *and their families*. Monographs of the Society for Research in Child Development, *57*: (6, Serial no. 230).

Stern, D. (1985). *The Interpersonal World of the Infant.* NY: Basic Books.

Sumner, G. and Spietz, A. (1994). *NCAST Caregiver/Parent-Child Interaction Teaching Manual*. Seattle: NCAST Publications, University of Washington, School of Nursing.

Trevarthen, C., and Aitken, K.J. (2001). Infant intersubjectivity: Research, theory and clinical applications. *Journal of Child Psychology and Psychiatry, 42*: 3–48.

Tronick, E. (1989). Emotions and emotional communication in infants. *American Psychologist, 44*: 112–119.

U.S. Department of Education, National Center for Education Statistics (2004). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Design and Operations Report for the Nine-Month Data Collection* (Unpublished report). Washington, D.C.: Author.

Walker-Andrews, A.S. (1998). Emotions and social development: Infants' recognition of emotions in others. *Pediatrics, 102* (5) Supplement: 1268–1271.

Zimowski, M.F., Muraki, E., Mislevy, R.J., and Darrell Bock, R. (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 Bayley Short Form, Research Edition (BSF-R) mental scale and motor scale items, 9-month data collection: 2001–02

Item	X1MTLTSC	X1MTLSCL	X1MTLSSE	X1MTL1	X1MTL2	X1MTL3	X1MTL4	X1MTL5	X1MTRTSC
X1MTLTSC	1.00	0.28*	-0.28*	0.35*	0.41*	0.28*	0.12*	0.05*	0.30*
X1MTLSCL	0.28*	1.00	-0.05*	0.53*	0.74*	0.98*	0.89*	0.73*	0.17*
X1MTLSSE	-0.28	-0.05*	1.00	0.14*	-0.03*	-0.19*	0.09*	0.32*	-0.05*
X1MTL1	0.35*	0.53*	0.14*	1.00	0.84*	0.51*	0.26*	0.16*	0.15*
X1MTL2	0.41*	0.74*	-0.03*	0.84*	1.00	0.74*	0.40*	0.25*	0.15*
X1MTL3	0.28*	0.98*	-0.19*	0.51*	0.74*	1.00	0.83*	0.60*	0.15*
X1MTL4	0.12*	0.89*	0.09*	0.26*	0.40*	0.83*	1.00	0.93*	0.16*
X1MTL5	0.053*	0.73*	0.32*	0.16*	0.25*	0.60*	0.92*	1.00	0.14*
X1MTRTSC	0.30*	0.17*	-0.05*	0.15*	0.15*	0.15*	0.16*	0.14*	1.00

Item	X1MTRSCL	X1MTRSSE	X1MTR1	X1MTR2	X1MTR3	X1MTR4	X1MTR5	X1MTLMOD	X1MTLSKP
X1MTLTSC	0.06*	-0.18*	0.03*	0.05*	0.02	-0.084*	-0.14*	-0.01	-0.04*
X1MTLSCL	0.72*	0.22*	0.54*	0.46*	0.55*	0.72*	0.63*	-0.01	-0.03*
X1MTLSSE	0.02*	0.27*	-0.01	-0.01	-0.06*	-0.44*	0.17*	-0.01	0.01
X1MTL1	0.36*	-0.07*	0.46*	0.45*	0.37*	0.28*	0.17*	-0.02*	-0.03*
X1MTL2	0.48	-0.08*	0.52*	0.49*	0.49*	0.43*	0.26*	-0.02*	-0.03*
X1MTL3	0.72	0.15*	0.55*	0.46*	0.56*	0.72*	0.57*	-0.01	-0.03*
X1MTL4	0.68*	0.41*	0.41*	0.32*	0.42*	0.69*	0.73*	-0.01	-0.02*
X1MTL5	0.56*	0.48*	0.29*	0.22*	0.30*	0.54*	0.69*	0.00	-0.01
X1MTRTSC	0.64*	-0.06*	0.63*	0.61*	0.64*	0.54*	0.42*	-0.03*	-0.05*

Item	X1MTRMOD	X1MTRSKP	X1CHLENG	X1CHWGHT	X1CHMUAC	X1CHCRFM	X1NCATTS	X1NCATTC	X1NCATTP
X1MTLTSC	0.01	-0.01	-0.17*	-0.10*	0.02	0.02*	0.07*	0.03*	0.07*
X1MTLSCL	0.01	-0.02*	0.51*	0.27*	0.06*	-0.05*	0.16*	0.06*	0.17*
X1MTLSSE	0.00	-0.01	0.05*	0.03*	0.03*	0.01	-0.04*	-0.02*	-0.04*
X1MTL1	-0.02*	-0.06*	0.22*	0.13*	0.05*	-0.10*	0.10*	0.06*	0.09*
X1MTL2	-0.01	-0.04*	0.32*	0.18*	0.05*	-0.09	0.14*	0.07*	0.14*
X1MTL3	0.01	-0.02*	0.50*	0.27*	0.05*	-0.05*	0.17*	0.061*	0.18*
X1MTL4	0.01	0.00	0.49*	0.26*	0.06*	-0.03*	0.13*	0.03	0.15*
X1MTL5	0.01	0.01	0.41*	0.21*	0.06*	-0.02*	0.10*	0.02*	0.11*
X1MTRTSC	-0.01	-0.02*	0.07*	0.03*	0.02	-0.04*	-0.01	-0.03*	0.00

* P < .05.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Nine-Month Data Collection, 2001–02.

This page is intentionally left blank.