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

Preschool—Kindergarten 2007
Psychometric Report

APRIL 2010

Michelle Najarian
Educational Testing Service

Kyle Snow
Jean Lennon
Susan Kinsey
RTI International

Gail Mulligan
Project Officer
National Center for Education Statistics

NCES 2010-009
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Statistics
Stuart Kerachsky
Deputy Commissioner

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

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

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

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

April 2010

The NCES World Wide Web Home Page address is http://nces.ed.gov.
The NCES World Wide Web Publications and Products address is http://nces.ed.gov/pubsearch.

Suggested Citation

Najarian, M., Snow, K., Lennon, J., and Kinsey, S. (2010). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Preschool–Kindergarten 2007 Psychometric Report* (NCES 2010-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

For ordering information on this report, write to

ED Pubs
U.S. Department of Education
P.O. Box 22207
Alexandria, VA 22304

or call toll free 1-877-4-ED-Pubs or order online at http://www.EDPubs.gov.

Content Contact

Gail M. Mulligan
(202) 502-7491
gail.mulligan@ed.gov

Acknowledgments

Since its inception, the design and implementation of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) has benefited from the contributions of many individuals and agencies. Although it is not possible to name all the individuals who have made significant contributions to this study, we would like to recognize their efforts in the development and implementation phases of the ECLS-B.

Finally, a special thank you to Kendra Chandler Webb, who, at the age of 9, designed the ECLS logo.

Contents

List of Tables

Table Page

List of Figures

Chapter 1
Introduction

This report describes the design, construction, implementation, quality control, and psychometric characteristics of the child assessment instruments used to measure developmental outcomes for young children participating in the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) during their transition into kindergarten. The focus of this volume is the final three waves of data collection: the preschool wave (2005–06); the kindergarten 2006 wave (2006–07); and the kindergarten 2007 wave (2007–08).[1] These waves were designed to examine child development during the year prior to kindergarten entry (i.e., the preschool wave) and at about the time of enrollment in kindergarten. This chapter provides a brief overview of the ECLS-B study; a discussion of the sample at each wave of data collection (section 1.1); an overview of the instrumentation (section 1.2); a description of the members and contributions of the technical review panel, who advised on the development of the assessments (section 1.3); and an overview of the contents of this report (section 1.4).[2]

The ECLS-B is a multisource, multimethod study that focuses on the early home and educational experiences of children from infancy to kindergarten entry. The study's primary sponsor is the National Center for Education Statistics (NCES) within the Institute of Education Sciences (IES) of the U.S. Department of Education. Cosponsors from the U.S. Department of Education include the National Center for Special Education Research (NCSER) of IES, the Office of Special Education Programs (OSEP), and the Office of Indian Education (OIE). Cosponsors from the U.S. Department of Health and Human Services include the Centers for Disease Control and Prevention (CDC); the National Center for Health Statistics (NCHS); the National Institute of Child Health and Human Development (NICHD); the Office of Behavioral and Social Sciences Research (OBSSR); the National Institute on Deafness and Other Communication Disorders (NIDCD); the National Institute of Mental Health (NIMH); the National Institute of Nursing Research (NINR); the National Institute on Aging (NIA); the National Center on Minority Health and Health Disparities (NCMHD); the Office of Planning, Research and Evaluation (OPRE) and the Child Care Bureau (CCB) in the Administration for Children and Families (ACF); the Office of the Assistant Secretary for Planning and Evaluation (ASPE); the Maternal and Child Health Bureau (MCHB); and the Office of Minority Health (OMH). The Economic Research Service (ERS) of the U.S. Department of Agriculture also cosponsors the study. Westat, a social science research firm, conducted the first two waves of the

[1] Psychometric reports have already been published for prior waves of data collection. See the ECLS-B 9-Month Psychometric Report (Andreassen, Fletcher, and West 2005) and the ECLS-B 2-Year Psychometric Report (Andreassen and Fletcher 2007). These are available from the U.S. Department of Education, National Center for Education Statistics website (http://nces.ed.gov/pubsearch).

[2] A list of abbreviations is included in appendix A.

study for NCES. RTI International conducted the preschool, kindergarten 2006, and kindergarten 2007 waves of the study.

The ECLS-B has followed a nationally representative cohort of children born in the United States in 2001 from birth through kindergarten entry. The study was designed as part of a longitudinal studies program comprising two cohorts: a birth cohort in the ECLS-B and a kindergarten cohort in the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K). The birth cohort study focuses on those characteristics of children and their families, including children's early learning, care, and health experiences, that may be associated with children's early development and kindergarten readiness. The kindergarten cohort follows a national sample of children who were in kindergarten during the 1998–99 school year, from kindergarten through eighth grade. The ECLS-K was designed to measure associations between home and school learning environments as children enter school and their academic achievement and experiences as they progressed through school. The two cohorts (ECLS-B and ECLS-K) were designed to overlap with respect to time points (i.e., both collect data about children during kindergarten) and with respect to measures (e.g., similar direct child assessment and parent interview constructs). Each study was designed to allow for longitudinal and cross-sectional analyses within each cohort, and the design of the ECLS-B was informed by that of the ECLS-K. This overlapping cohort feature was designed to improve understanding of where, when, and how differences observed in kindergarten in children's cognitive, social, and physical development may have emerged (see West, Denton, and Germino-Hausken 2000).

The parents of about 10,700 children born in 2001 participated in the first wave, or 9-month wave, of the ECLS-B study when the children were approximately 9 months of age. Child assessments were conducted with about 10,200 of these children between October 2001 and December 2002, to correspond closely with their reaching 9 months of age. The second or 2-year wave of the study was conducted in 2003, when the children were approximately 2 years old. The parents of about 9,850 children participated in the 2-year wave, and child assessments were conducted with about 9,200[3] of these children. Because the sampled children were born between January and December 2001, and it was a study goal to conduct the child assessments as close to the child's second birthday as possible, the second wave of data was collected on a rolling basis between January 2003 and December 2003. However, as the study children aged, the focus of the study shifted toward obtaining indicators of school readiness. Therefore, the remaining data waves—preschool, kindergarten 2006, and kindergarten 2007—were conducted to coincide with

[3] This number corresponds to the number of children who have a parent interview at the 2-year wave of data collection and have at least one of the following types of assessment data: both height and weight, a BSF-R mental score, or a BSF-R motor score. There are several additional cases on the data file that have at least one of these types of assessment data but no parent interview information.

the academic year to better assess skills related to school readiness. Section 1.1 has a description of the samples included in the preschool and kindergarten waves.[4]

The central goal of the ECLS-B is to provide a comprehensive and reliable set of data that may be used to better understand and describe children's early development; their heath care, nutrition, and physical well-being; their home learning experiences; their experiences in early care and education programs and at kindergarten entry; and how their early experiences relate to their later development, learning, and experiences in school.

Several aspects of the ECLS-B design are unique compared with other studies on early childhood development and growth. These aspects are specially suited to the goals of describing the range of developmental experiences for children between birth and kindergarten entry. These include oversampling of specific groups of children (e.g., American Indian and Alaska Native infants, Asian and Pacific Islander infants, Chinese American infants,[5] very low birth weight and moderately low birth weight infants, and twins); collecting information directly from children's mothers, fathers, early care and education providers (ECEPs), wrap-around (i.e., before- and after-school) early care and education providers (WECEPs), and teachers; video recordings of parent-child interactions; audiotaping of children's natural language expression; and observing child care settings serving the ECLS-B sampled children.[6] These features enrich the study design and provide in-depth descriptions of children's early home and care experiences and their later development and kindergarten experiences. They also support unique analyses of outcomes for groups of children whose proportional representation in the overall population would not otherwise allow precision in statistical estimates.

The ECLS-B may be used for both descriptive and analytic purposes. It provides descriptive data on children's health status at birth and later; their experiences in home, nonparental care, and school environments; and their development and growth up to kindergarten entry.[7] The ECLS-B provides a rich source of data that enables researchers to analyze how a wide range of family, nonparental care, school, community, and individual factors are associated with children's early experiences, development, and school readiness; to explore school readiness and the relationship between early care and education experiences and later school experiences; and to record children's cognitive, socioemotional, and physical growth from infancy up to the first year of formal schooling.

[4] As discussed in chapter 2, much of the instrumentation was the same for the two kindergarten waves of data collection. As a result, this report uses the phrase "kindergarten waves" unless specific differences between the two waves need to be highlighted, in which case the waves are specified (i.e., kindergarten 2006 or kindergarten 2007).

[5] Chinese American infants were oversampled separately from the oversample of Asian and Pacific Islander infants to allow precision in comparisons between the full race/ethnicity group and its largest represented subgroup.

[6] Although a parent interview and direct child assessment were included for all waves of data collection, other components of the study were included only during specific waves or for specific subsamples of children. See section 1.2 for an overview of study components at each wave of data collection.

[7] Although the focus is on all study children as they enter kindergarten, the kindergarten 2007 wave also includes children who were first in kindergarten in 2006 and were repeating kindergarten in 2007. This is described further in section 1.1.

1.1 The Preschool to Kindergarten Child Sample

As noted above, the ECLS-B is an age-based cohort. Participating families were visited for the first two waves of data collection around specified child ages (9 months and 2 years); that is, the goal was to visit the children's homes as close as possible to the date on which they turned the target age. However, starting with the preschool wave, the design shifted from one that was strictly age-based to one for which the goal was to collect data from families and children at a particular level of schooling. Thus, the preschool wave was intended to capture data during the year before most study children would be expected to enter kindergarten. As a result, the timing of data collection was designed to coincide with the academic year, from the fall of 2005 to the spring of 2006.

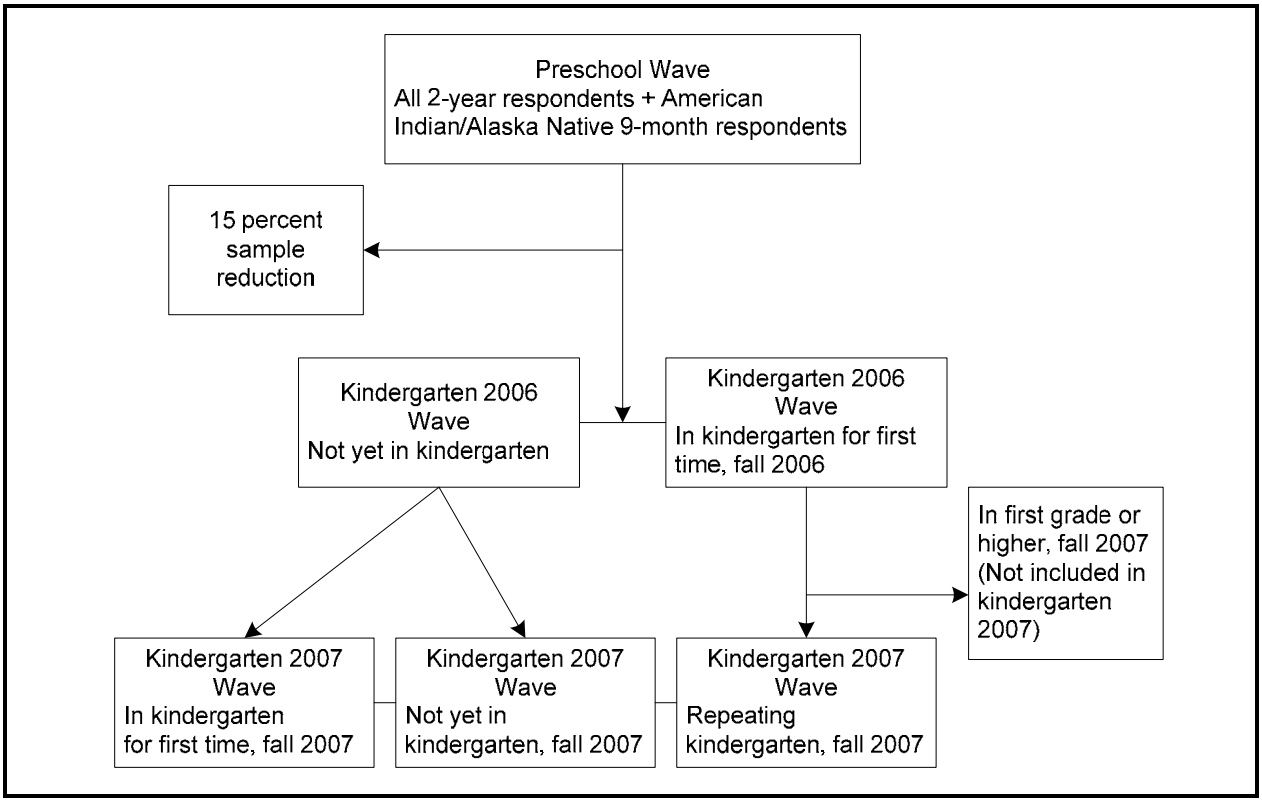
The kindergarten 2006 wave, conducted from fall 2006 through early spring 2007, was designed to collect data from study children when they were first in kindergarten. However, because the ECLS-B sample is an age-based cohort, children were age-eligible for kindergarten in two different academic years. The majority of children became eligible for kindergarten during fall 2006 (i.e., the kindergarten 2006 wave), while the other children started kindergarten during fall 2007 (i.e., the kindergarten 2007 wave). The 2006 collection included all children, whether or not they had entered kindergarten; approximately 75 percent of children were in kindergarten or higher for this wave.[8] The kindergarten 2007 wave was conducted between fall 2007 and spring 2008 and included the children who did not enter kindergarten in fall 2006, some of whom had not been age-eligible and some of whom had a delayed entry. The 2007 collection also included the approximately 5 percent of children who were kindergartners in fall 2006 and were identified as repeating kindergarten in fall 2007. This sampling design is summarized in figure 1.

The ECLS-B is representative of children born in the United States in 2001. It is not representative of any cross-section of children thereafter. For example, wave 3 (preschool) is not representative of children in preschool. Thus, while the preschool wave was intended to describe children's knowledge, skills, and abilities during the year prior to kindergarten entry, not all children in the preschool wave sample were in a formal preschool program and not all children started kindergarten the following year. The preschool data represent children born in the United States during 2001 when they were about 4 years old, the majority of whom were expected to start kindergarten the following year. Similarly, waves 4 and 5 (kindergarten 2006 and kindergarten 2007) are not representative of children in kindergarten. The ECLS-B was designed to provide information on a birth cohort, and the follow-up waves of collection (i.e., waves 2 through 5) are meant to provide information on the birth cohort's experiences leading up to and including kindergarten entry. So, for example, while the wave 4 and wave 5 data are not

[8] At the kindergarten 2006 wave, the ECLS-B sample was reduced by 15 percent for budgetary reasons. The remaining 85 percent were eligible for follow-up. The subsample was selected using a stratified random sample, with strata defined by the original race/ethnicity, birth weight, and plurality subgroups of the population. The subsample was allocated disproportionately to the subgroups to preserve the precision of estimates for the smaller subgroups. In particular, all children who were sampled in the American Indian/Alaska Native and Chinese American subgroups were retained in the subsample.

representative of kindergartners, they are representative of the children born in 2001 when they entered kindergarten.

Figure 1. Summary of the ECLS-B sample, preschool through kindergarten 2007 data collections: 2005–06, 2006–07, 2007–08



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Selected demographic information is provided in table 1 describing the child sample at the preschool, kindergarten 2006, and kindergarten 2007 waves. For the preschool wave, parents of about 8,950 children were interviewed, and child assessments were conducted with about 8,750 children.[9] Data were collected from parents for approximately 7,000 children during the kindergarten 2006 wave and approximately 1,900 children during the kindergarten 2007 wave. Of the 1,900 children participating in the 2007 data collection, approximately 1,550 were first-time kindergartners (nonrepeaters) and nearly 200 children were repeating kindergarten in fall 2007.[10] The remaining children (approximately 200) cannot be categorized as first-time kindergartners or kindergarten repeaters, because they were either enrolled in multigrade or ungraded classrooms, enrolled in the second year of a two-year kindergarten program, homeschooled, or not in school during the fall of 2007.

[9] During the processing of the kindergarten 2006 data, the definition of a child-level complete for the preschool wave was revised to be consistent with the kindergarten 2006 definition. This resulted in 10 fewer preschool child-level completes than are reported in the ECLS-B Preschool Data File User's Manual (Snow et al. 2007).

[10] These numbers are based on the XKWHENK and X5RPTR composites on the data file. For information on the construction of these variables please see chapter 10 in the ECLS-B kindergarten 2006 and 2007 data file user's manual (Snow et al. 2009).

Table 1. Selected demographic characteristics of the ECLS-B child sample during the preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Characteristic	Preschool	Kindergarten 2006	Kindergarten 2007
Mean child age (months)	52.46	64.71	74.22
Child sex			
Male	51.2	51.2	54.7
Female	48.8	48.8	45.3
Child race/ethnicity			
White, non-Hispanic	53.8	53.8	55.9
Black, non-Hispanic	13.8	13.9	14.8
Hispanic, race specified	16.9	16.9	14.8
Hispanic, no race specified	8.2	8.2	8.4
Asian, non-Hispanic	2.6	2.6	1.7
Native Hawaiian/Pacific Islander, non-Hispanic	0.2	0.2	0.1
American Indian/Alaska Native, non-Hispanic	0.5	0.5	0.6
More than one race specified, non-Hispanic	4.0	3.9	3.8
Poverty status (at time of collection)[1]			
Below poverty threshold	24.8	24.3	25.8
At or above poverty threshold	75.2	75.7	74.2
Birth weight			
Normal (2,500 grams or more)	92.5	92.5	91.9
Moderately low (≥ 1,500 grams and < 2,500 grams)	6.2	6.2	6.5
Very low (less than 1,500 grams)	1.3	1.3	1.6

[1]The preschool estimates indicate poverty status for the preschool wave, the kindergarten 2006 estimates indicate poverty status for the kindergarten 2006 wave, and the kindergarten 2007 estimates indicate poverty status for the kindergarten 2007 wave.

NOTE: Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data.

Numbers shown are percents, unless otherwise indicated.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

1.2 Data Collection Instruments and Administration of Assessments

The preschool, kindergarten 2006, and kindergarten 2007 data collections included in-person home interviews with primary caregivers (most often the child's mother) using computer-assisted personal interviewing technology, direct assessments of children's cognitive and motor abilities, and measurement of the child's physical growth. The data collection instruments used during each of these waves are shown in table 2. During the preschool wave, the child and parent also participated in a semistructured play task (the Two Bags Task) that was recorded for later coding. At this wave the interviewer asked for the name, address, and phone number of the early care and education provider (if any) with whom the child spent the most time on a weekly basis. Telephone interviews were conducted with the child's primary ECEP. For a subsample of the child care providers/teachers, an observation of the early care and education setting also was

conducted. In addition, during the preschool wave, self-administered questionnaires were completed by the resident father (i.e., the partner or spouse of the parent respondent, who could be the child's biological father, adoptive father, stepfather, foster father, or other type of father figure).

Table 2. Assessments and instruments used in the ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

ECLS-B data sources and instruments	Preschool	Kindergarten 2006	Kindergarten 2007
Home visit			
Direct child assessments (computer-assisted interview, paper-and-pencil interview, easel, audiotapes, and physical measurements)	X	X	X
Parent interview			
Parent CAPI Instrument	X	X	X
Parent Self-Administered Questionnaire (PSAQ) via ACASI	X	X	X
Resident Father Self-Administered Questionnaire (RFSAQ)	X	—	—
Child-Parent Observations via semi-structured play task (Two Bags Task)	X	—	—
Early care and education provider (ECEP) telephone interview (children not in kindergarten or above)	X	X	—
Child Care Observation (CCO) for a subset of the children interviewed	X	—	—
Wrap-around early care and education provider (WECEP) telephone interview (children in kindergarten)	—	X	X
Teacher Self-Administered Questionnaire (TSAQ) (children enrolled in kindergarten or above)	—	X	X
Extant Supplemental Sources			
Common Core of Data (CCD) data file	—	X	X
Private School Universe Survey (PSS) data file	—	X	X

NOTE: ACASI = audio computer-assisted self-interviewing. CAPI = computer-assisted personal interviewing.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

During the parent interview for the kindergarten 2006 wave, the interviewer asked about the child's kindergarten enrollment and the amount of time the child spent in nonparental care.[11] For children not in kindergarten or higher but who regularly spent at least 1 hour per week in a nonparental setting, telephone interviews were conducted with the child's primary ECEP. For children enrolled in school (kindergarten or beyond), their teachers were asked to complete a self-administered questionnaire. For children in kindergarten and receiving at least 5 hours per week of regularly scheduled nonparental care, interviews also were conducted with the child's WECEPs. The kindergarten 2007 wave was comparable to the kindergarten 2006 wave with respect to the components fielded, although there was no ECEP interview.

Prior waves of the ECLS-B assessed children using the Bayley Short Form—Research Edition (BSF-R; see Andreassen, Fletcher and West 2005). However, as described in chapter 2 of this report, new assessments were developed for the preschool and kindergarten waves that were more age-appropriate for the study children. During the preschool and kindergarten waves, the direct child assessment was a 1-hour individualized evaluation of children's development that included a direct cognitive assessment, an assessment of gross and fine motor abilities, and measurements of physical growth (i.e., height, weight, middle upper arm circumference, and, for very low birth weight children, head circumference) (table 3).

Table 3. Components and substantive domains covered in the ECLS-B preschool and kindergarten direct child assessments: 2005–06, 2006–07, and 2007–08

Direct child assessment component	Domain coverage
Adaptive tests in early reading, mathematics, and color knowledge[1]	Cognitive (mental)
Physical measurements (height, weight, middle upper arm circumference, head circumference)[2]	Physical growth and development
Motor skills (building blocks, drawing geometric figures, jumping, balancing, hopping, skipping, walking backward, bean bag catch)	Hand-eye coordination, general muscle growth, and physical skill

[1] Color knowledge task was administered at preschool only.

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

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

The data collection questionnaires used during the preschool and kindergarten waves are available in appendix A of the ECLS-B DVD.[12] However, the direct child assessment and certain parent, ECEP, and WECEP items are not available on the DVD because they are copyright-protected materials and agreements with the publishers restrict their distribution. These assessments may be requested from NCES once publisher permission has been obtained. See "Guidelines for the Release and Use of ECLS-B Copyrighted Measures" at http://nces.ed.gov/ecls/pdf/Birth/ECLSB_Copyright_Guidelines.pdf.

[11] At the kindergarten 2006 data collection, 75.2 percent of participating children were in kindergarten or higher based on parent reports. Children reported by the parent respondent as being enrolled in kindergarten or higher were also eligible for the Teacher Self-Administered Questionnaire and a WECEP interview. Children not reported by parents to be enrolled in kindergarten were eligible for an ECEP interview. If a child was homeschooled, the case was not eligible for a teacher survey but may have been eligible for a WECEP interview.

[12] For information on gaining access to the ECLS-B data, please refer to http://nces.ed.gov/ecls/birth.asp.

The direct child assessments were administered during the home visit by a staff of trained field interviewers (FIs). The FIs were trained to begin the home visit with the parent interview to provide the child with the opportunity to become accustomed to the FI and so the FI could build rapport with the child before beginning the child assessment. The design allowed for the FI to halt the parent interview to conduct the child assessments, and then complete the parent interview, if that seemed like the best way to incorporate the assessment into the home visit for a particular household.

During the preschool and kindergarten waves of data collection, the direct child assessment procedures were generally the same. The direct child cognitive assessment began with the parent being asked a series of items designed to determine whether the child had any physical limitations that would preclude him or her being administered the assessment, as well as the child's understanding of English and Spanish. Parent responses to these items drove the administration of the direct child assessment.

Because the direct cognitive assessment relied on auditory and visual stimuli, children who required Braille or sign language were not administered the cognitive component of the direct child assessment. Likewise, because some fine and gross motor assessments required an ability to see, children requiring Braille were not administered the fine or gross motor items requiring sight. Children requiring sign language could complete both the gross and fine motor portions of the assessment. Children in wheelchairs could complete the direct cognitive and fine motor components, as well as the bean bag catch gross motor item, but were not administered the other motor items. FIs also were allowed to skip individual items if a child could not complete them because of a specific, perhaps temporary, limitation. For example, a child with his or her arm in a cast may not have been administered all of the motor items.

The direct cognitive assessment was available only in English or Spanish.[13] Children were routed to an assessment conducted in English in all cases where the child answered at least one of the nonpractice language routing items correctly (regardless of parental indication of English or Spanish proficiency). Children were routed to direct assessments conducted in Spanish if they did not correctly answer at least one of the nonpractice English language routing items and the parent indicated that the child understood Spanish. If the child did not respond correctly to any of the English language routing items, and the parent did not indicate that the child spoke Spanish, the direct cognitive assessment was ended. For these cases, the fine and gross motor assessments were administered using a translator (if available). Finally, all children participated in the physical measures component of the child assessment, although height and weight were not obtained for children in wheelchairs.

[13] The language of assessment was determined by both child performance on a set of language routing items and parent report of child language. Once determined, the language of assessment was constant—all components were conducted in the determined language. Please note that there are no Spanish child assessment scores on the file. Too few children took the Spanish cognitive assessment to meet sample size requirements in IRT analysis, so it was not scored.

1.3 Technical Review Panel

Since the inception of the study, design of the content of the child assessments for the ECLS-B has been guided by a Technical Review Panel (TRP) composed of nationally recognized experts in a variety of areas, including children's cognitive, language, and socioemotional development; health; family influences; fathers; child care; and community and other influences on development. Members of the TRP reviewed questionnaires, assessment instrument content, and draft assessment score plans. An important responsibility of the TRP was to ensure that the plans for conducting the ECLS-B were well thought out and complete, and this responsibility required a broad range of expertise. Members of the TRP and their areas of expertise are provided in table 4.

Table 4. The ECLS-B Technical Review Panel members, with affiliation and area of expertise: 1998–2008

Technical Review Panel member name	Affiliation	Area of expertise
Martha Abbott-Shim	Quality Counts, Inc. Quality Research Center	Quality of child care
Emily Arcia	Diversity Compliance Miami-Dade County Public Schools	Latino family issues, attention deficit hyperactivity disorder
Kathryn Barnard[1]	University of Washington School of Nursing	Early parent-infant relationships and effects on development
Susan Bredekamp	Child Development Associates Council for Early Childhood Recognition	School readiness, policy issues
Martha J. Cox	University of North Carolina at Chapel Hill FPG Child Development Center Department of Psychology	Parent-infant relationships and children's security of attachment
Susan Fowler	University of Illinois at Urbana-Champaign College of Education	Children's language acquisition and use through testing and observation
Thomas Jordan[2]	University of Missouri–St. Louis	Longitudinal studies of children
Milton Kotelchuck	Department of Maternal and Child Health Boston University School of Public Health	Pediatrics and child health policy
Kristin Moore	Child Trends	Father involvement
Barbara Alexander Pan	Harvard Graduate School of Education	Conversation and language between parents and children

See notes at end of table.

Table 4. The ECLS-B Technical Review Panel members, with affiliation and area of expertise: 1998–2008—Continued

Technical Review Panel member name	Affiliation	Area of expertise
Elizabeth Peters	Cornell University Department of Policy Analysis and Management	Economics of the family including child support, child care, marriage, and divorce
Suzanne Randolph	University of Maryland Department of Family Studies	Child development among African American families, parent-child interaction
Aline Sayer	Radcliffe Institute for Advanced Study Harvard University	Multilevel modeling, growth curve analysis
Heidi Schweingruber	Board on Science Education Center for Education The National Research Council	Early childhood mathematical and science cognition, early childhood programs and pedagogy
Susan Spieker	University of Washington	Infant and child socioemotional development, child care
Brian Vaughn	Human Development and Family Studies	Attachment research, social and personality development during infancy and childhood, and development of social competence
Barbara Wasik	The Johns Hopkins University	Direct assessment and testing of children's language
Barry Zuckerman	Boston University School of Medicine Department of Pediatrics	Low birth weight children, general child health and development

[1] Panel member for waves prior to the preschool wave.
[2] Panel member for waves prior to the kindergarten 2006 wave.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month (2001–02), 2-year (2003–04), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

1.4 Contents of This Report

This volume provides technical details about the development, design, and psychometric characteristics of the direct and indirect child assessments used during the preschool and kindergarten waves of data collection. Chapter 2 provides details about the design of the direct child cognitive assessment battery for the ECLS-B preschool and kindergarten waves. Chapter 3 provides an overview of the analytic methodology used to develop the direct child cognitive assessments. Chapter 3 also describes the methodology used to develop a longitudinal scale for the assessments from preschool to kindergarten, including analysis of common item functioning and recalibration of scores. Chapter 4 provides detailed analyses of the psychometric characteristics of the direct child cognitive assessments during each of the three waves of data collection, including item analysis for reliability and differential functioning and score statistics. Chapter 5 provides information on child physical measurements and the psychomotor

assessments used in the preschool and kindergarten waves. Chapter 6 provides information on measures of parent-child interaction, using the Two Bags Task and Reading Aloud Profile–Together (RAPT) scoring systems. (Please also refer to chapter 3 of the ECLS-B Kindergarten 2006 and 2007 Data File User's Manual [Snow et al. 2009] for more information on the coding scheme used to analyze the parent-child interactions in the Two Bags Task and RAPT.) Chapter 7 describes the indirect child measures obtained from parents, teachers, ECEPs, and WECEPs across the three waves of data collection. Chapter 8 provides information on the coding of language and fine motor items. Details about the ECLS-B direct child cognitive assessment IRT parameters are provided in appendix B. Additional information about the ECLS program may be found at http://nces.ed.gov/ecls.

Chapter 2
Design and Development of the Direct Cognitive Assessment

The ECLS-B directly assessed children's early reading and mathematics skills during the preschool, kindergarten 2006, and kindergarten 2007 data waves. This chapter documents the design of the direct cognitive assessment instruments used in the ECLS-B preschool and kindergarten (2006 and 2007) data collection waves. Section 2.1 presents the guidelines for designing the cognitive assessments and an overview of the assessment framework. Section 2.2 outlines the development of the preschool wave assessment. Section 2.3 provides information on the development of the assessment for the kindergarten waves.

2.1 Designing the ECLS-B Preschool and Kindergarten Direct Cognitive Assessments

During the 9-month and 2-year data waves the Bayley Short Form-Research Edition (BSF-R)[14] was used to directly assess children's developmental status in terms of their mental and psychomotor skills. However, once study children were 4 years of age and nearing formal school entry, it was deemed important to assess children's cognitive development with respect to those skills more directly related to school readiness. Thus, the preschool and kindergarten data waves shifted to a cognitive assessment that evaluated children's early reading and mathematics skills. New assessment instruments were developed that were age-appropriate, targeted the aforementioned skills, and allowed for the analysis of cognitive gains over time (i.e., from preschool to kindergarten). In developing the preschool and kindergarten direct assessment batteries, certain considerations guided the design of the instruments:

The design of the child assessment took into account the need for reliable, standardized administration and scoring and appropriateness for a home setting. While ECLS-B field staff had expertise in administering survey instruments, most did not have extensive knowledge of child development and assessment. It was necessary to train a large number of field staff (approximately 450 people for the preschool collection, 250 for the kindergarten 2006 data collection, and 80 for the kindergarten 2007 wave) to administer the preschool and kindergarten cognitive assessments in a standardized way. To achieve reliable, standardized administration, the assessment instruments needed to have very clear administration guidelines for which only a basic understanding of child development was required. Also, measures had to be appropriate for use in a home setting and have scoring procedures that were standardized and fairly easy for field staff to use. Additionally, the ECLS-B design featured many different tasks, each requiring comprehensive training and use of special skills. For example, while in the home, a field interviewer completed a

[14] The BSF-R is a modified version of the Bayley Scales of Infant Development–II, created specifically for the ECLS-B with the publisher's permission.

number of discrete tasks (e.g., conducted the parent interview, assessed the child, collected child care provider and teacher information and, during the preschool wave, videotaped parent-child interactions). While separately no one task was difficult, the total data collection protocol was complex, so it was essential that the child assessment be fairly simple to administer.

The ECLS-B child assessment was designed to accommodate children with varying needs and abilities. By design, children in the ECLS-B sample had varying levels of ability and were living in all types of home situations. Additionally, variability in young children's abilities was to be expected during this period of rapid development in early childhood. Thus, the ECLS-B kindergarten assessment needed to cover a wide range of abilities across the multiple domains. Untimed and one-on-one administration, with the possibility of breaks during administration if necessary, made it so that the assessment could accommodate children with varying needs and levels of ability.

The ECLS-B needed to maximize information gathered in a short time frame. One way to maximize information gathered in a short time frame is to carefully consider the mode of administration (e.g., hard copy versus computer-assisted). In the 9-month and 2-year waves of the ECLS-B, child assessment data were recorded manually in a hard-copy booklet. In the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), the direct child cognitive assessment was computer-assisted, with field staff asking children questions based on materials presented on easels and recording children's answers directly into a laptop computer. This mode of administration facilitates the use of adaptive tests because the computer automatically makes decisions about skip patterns and choices of test forms without the need for the assessor to keep track of children's performance during the test. Also, individually administered adaptive tests can reduce administration time and increase measurement accuracy by selecting test items that are most appropriate for each child's ability level. The success of this approach in the ECLS-K led to its use in the ECLS-B preschool and kindergarten waves.

The ECLS-B needed to be as inclusive as possible of children with limited fluency in English. Many of the children in the ECLS-B sample came from homes where English was not the primary language spoken. The ECLS-B preschool and kindergarten wave assessments adapted the approach that was used for 5- and 6-year-olds in the ECLS-K, whereby Spanish translations of the assessments were used for Spanish-speaking children who did not pass an English fluency screening measure. For the preschool and kindergarten waves of the ECLS-B, a Spanish-language version of the PreLAS (Preschool Language Assessment Scale; Duncan & De Avila 1998) was given to assess the language skills of non-English speakers who spoke Spanish. A Spanish Peabody Picture Vocabulary Test (the Test de Vocabulario en Imagenes Peabody [TVIP; Dunn et al 1986]) also was administered, as well as a Spanish translation of the mathematics assessment.

The objectives of the preschool and kindergarten wave child assessments remained similar to the objectives of the 9-month and 2-year assessments[15]—to provide measures of children's development and functioning across their first years of life. However, the assessment design used in the preschool and kindergarten waves differed from the 9-month and 2-year design in several important ways due to the aging of the cohort and the increased salience of academically oriented skills.

The assessment window was changed depending on the month of the child's birthday to coincide with the academic calendar. Thus, the preschool wave began in fall 2005, the kindergarten 2006 wave began in fall 2006, and the kindergarten 2007 wave began in fall 2007.

Starting with the preschool wave, adaptive tests in early reading, color knowledge, and mathematics replaced the Bayley Short Form-Research Edition (BSF-R). (Note that the color knowledge assessment was included only at preschool.)

A language screening test was used to determine whether children's English fluency was sufficient for the level of communication necessary for participation in the English direct cognitive assessments.

2.1.1 Description of the ECLS-B Preschool and Kindergarten Cognitive Assessment Framework

The ECLS-B approach of "breadth over depth" led to the development of instruments that broadly survey children's knowledge and skills across several domains. In choosing the constructs that were most important to include in the preschool and kindergarten assessment batteries, key developmental milestones at these ages were considered, as well as the knowledge and skills that are important for school readiness and early school success. Thus, the ECLS-B assessment framework combined a developmental age perspective with a focus on academic curriculum content.

Children's knowledge in three key cognitive domains was assessed in the preschool and kindergarten waves: early reading, mathematics, and color knowledge (preschool only). As in the 9-month and 2-year waves, the time allotted for the preschool and kindergarten direct cognitive assessment was approximately 35 minutes (10 additional minutes were allotted for assessing physical and motor skills, totaling 45 minutes of direct assessment time). Table 5 outlines the final content framework for the ECLS-B preschool and kindergarten cognitive assessment.

[15] For additional information on the 9-month and 2-year assessment designs and results, see the ECLS-B 9-Month Psychometric Report (Andreassen, Fletcher, and West 2005) and the ECLS-B 2-Year Psychometric Report (Andreassen and Fletcher 2007).

Table 5. Content framework for the ECLS-B cognitive assessment

Early reading	Mathematics	Color knowledge
Basic Skills	Number Sense, Properties, and Operations	Color Knowledge (Preschool only)
English language skills	Measurement	
Letter knowledge	Geometry and Spatial Sense	
Letter-sound knowledge	Data Analysis, Statistics, and Probability	
Print conventions	Patterns, Algebra, and Functions	
Word recognition		
Vocabulary		
Initial Understanding		
Developing Interpretation		
Demonstrating Critical Stance		

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

While the assessment design was intended to provide continuity across waves and sufficient item overlap to allow for the development of longitudinal scales for measuring early reading and mathematics knowledge and skills, the items included in each of the assessments were tailored to be appropriate to child developmental level. The final content of the early reading and mathematics assessments was guided by frameworks provided by Brush et al. (2003) for the preschool wave and was based upon the design of the ECLS-K kindergarten year assessment for the kindergarten waves. As a result, the distribution of items across skill areas differed between the preschool and kindergarten waves.

The constructs presented in table 5 related to **early reading** are the following:

Basic Skills, comprising several categories, including English-language skills and oral language, phonological awareness, letter and letter-sound knowledge, print conventions, and word recognition.

Vocabulary, which requires skills in both receptive and expressive language. Receptive language does not rely on oral language skills and thus may provide a more accurate measure of children's conceptual knowledge than measures of expressive language; consequently, both were assessed.

Initial Understanding, which requires those assessed to provide an initial impression or global understanding of context. Identifying the main point and the specific points used to construct that main point would be included in this category.

Developing Interpretation, which requires those assessed to extend their initial impressions to develop a more complete understanding of context. It involves the linking of information across parts, as well as focusing on specific information.

Demonstrating a Critical Stance, which requires those assessed to demonstrate an understanding of the story. Being asked questions about a story and replying with answers that are not explicitly part of the text passage would be included in this category.

During the development of the early reading assessment, consideration was given to the inclusion of items that would assess children's early writing skills. However, an assessment of

early writing skills is time and resource intensive and was considered not to be feasible to include in the preschool and kindergarten waves of collection due to time and budget constraints.

Table 6 lists the targeted number and percentage of items in the ECLS-B preschool and kindergarten early reading assessments. The ECLS-B preschool early reading assessment was composed of items in the basic skills and vocabulary categories. Because the preschool early reading assessment primarily included basic skills content items, the construct was further subdivided into more refined content areas for the preschool wave to reflect recommendations from Brush et al. (2003). The ECLS-B kindergarten early reading assessment included items in the basic skills category as well as items tapping the other constructs identified in table 5.

Table 6. ECLS-B preschool and kindergarten framework targets for early reading content area: 2005–06, 2006–07, and 2007–08

	Preschool		Kindergarten	
Content category	Number of items	Percent of items	Number of items	Percent of items
Total	81	100	72	100
Basic skill	65	80	40	55
English language skills/oral language	22	—	†	—
Phonological awareness	18	—	†	—
Letter and letter-sound knowledge	12	—	†	—
Print conventions	7	—	†	—
Word recognition	6	—	†	—
Vocabulary	16	20	8	11
Initial understanding	0	0	20	28
Developing interpretation	0	0	2	3
Demonstrating a critical stance	0	0	2	3
Personal reflection and response[1]	0	0	0	0

— Not available.

† Not applicable.

[1] Although 10 percent of time was targeted for Personal Reflection and Response items in the ECLS-K, these items were more appropriate for the higher difficulty assessment developed for use in first grade, and thus were not included in the ECLS-B kindergarten design.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

The primary mathematics constructs included in the ECLS-B assessments are the following (see also table 5):

Number Sense, Properties, and Operations, which refers to children's understanding of numbers, cardinality, ordinality, quantity, operations, and estimation, and their application. Number sequence and understanding of one-to-one correspondence, including the idea that the last number counted represents a characteristic of a set of objects, are key concepts. Number sense and counting lay the groundwork for later development of skills in addition and subtraction. More advanced children may be able to solve simple problems involving part-whole concepts and simple operations, such as adding to or taking away, for various set sizes. To assess number sense, the

ECLS-B assessment tapped children's understanding of the concept that numbers represent quantity and their understanding of number words and symbols.

Measurement, which includes understanding the attributes of objects (e.g., length and volume) and the ability to compare objects by their attributes.

Geometry and Spatial Sense, which requires the ability to analyze the characteristics and relationships of shapes and concepts of spatial relations and geometric reasoning. Skills included in this content area extend from simple identification of geometric shapes to transformations and combinations of those shapes.

Data Analysis, Statistics, and Probability, which includes the skills of collecting, organizing, reading, and representing data. Children were asked to describe patterns in the data or to make inferences or draw conclusions based on the data.

Patterns, Algebra, and Functions, which requires children to identify, duplicate, and extend patterns that may predict later algebraic thinking about the properties of items. All the items included in this category were pattern recognition items.

As with the early reading assessment, the design of the preschool mathematics assessment was guided by the frameworks provided by Brush et al. (2003) and the design of the ECLS-K kindergarten assessments. The frameworks included the same constructs, but Brush et al. (2003) included number sense, counting, and operations, as three, rather than one, constructs. Table 7 shows the framework recommendations for the ECLS-B preschool and kindergarten mathematics assessment by construct.

Table 7. ECLS-K and ECLS-B kindergarten framework targets and actual item counts by content area, mathematics: 1998–99, 2006–07, and 2007–08

	Preschool		Kindergarten	
Content categories	Number of items	Percent of items	Number of items	Percent of items
Total	43	100	65	100
Number sense, properties, and operations	32	74	49	75
Number sense	14	—	†	—
Counting (includes color bears)	12	—	†	—
Operations	6	—	†	—
Measurement	2	5	3	5
Geometry and spatial sense	6	14	2	3
Data analysis, statistics, and probability	0	0	5	8
Patterns, algebra, and functions	3	7	6	9

‒ Not available.

† Not applicable.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K), kindergarten data collection, 1998–99, and Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections.

In preschool and kindergarten, general knowledge can be conceptualized in terms of basic concept attainment. Basic concepts include understanding and using relational concepts for size, distance, time, and quantity (e.g., on, under, tall, thin, more). A very broad notion of general

knowledge could extend to include the understanding of colors, shapes, letters, and numbers. Many of these basic concepts were assessed as part of the ECLS-B assessment of children's early reading and mathematics. Therefore, the ECLS-B did not include a separate general knowledge test, with the exception of a measure of children's color knowledge that was appropriate for administration at the preschool wave. Color knowledge is defined as children's ability to recognize and name basic colors such as red, blue, yellow, green, white, and black.

2.2 Preschool Assessment

This section presents an overview of the steps taken to develop the ECLS-B preschool assessment, including item pool development (section 2.2.1), field test work (section 2.2.2), analyses of field test data (section 2.2.3), and assembly of the main study (section 2.2.4).

2.2.1 Preschool Item Pool Development

The American Institutes for Research (AIR) developed and recommended item pools for the ECLS-B preschool wave (Nathanson et al. 2003) based on the above described framework. A small pilot study was conducted in the fall of 2002; using information obtained in the pilot study, items were refined, and then a larger pilot study was carried out in 2003. Informed by results from the larger pilot study, AIR recommended items and constructs to be used in a full-scale field test conducted in the fall of 2004. Analysis of field test results guided the revision of items and the design and assembly of test forms for use in the preschool wave.

The larger 2003 pilot test contained items pertaining to three early reading constructs (language development, emergent literacy, and beginning reading) and six mathematics constructs (number sense, counting, operations, geometry, patterns, and measurement). Each child participating in the pilot test received either the early reading assessment (containing both language and literacy items) or the mathematics assessment. The pilot test largely consistent of 161 early reading and 83 mathematics items, taking an average of 1 hour to administer the items in one domain or the other.[16] The pilot test also contained practice items, one or two for each item presentation type (a total of 30 practice items in early reading and 31 in mathematics) to introduce children to the format of each task. To minimize burden and possible fatigue, children were encouraged to take short breaks when needed. To ensure standardization, the assessment items were administered following guidelines provided with the original source for the item. About 50 children, with an average age of 4.3 years, completed the early reading assessment. About 50 children, with an average age of 4.4 years, completed the mathematics assessment.

[16] At the start of the 2003 pilot test, the early reading assessment item pool contained 214 items and the mathematics assessment item pool had 118 items. The first few children in the pilot received the entire pool (12 children in reading; 7 children in math). It was quickly realized, however, that administration time for the items in a single domain (approximately 2 hours) was too long to be appropriate for preschool children. Therefore, the item pool for each domain was revised to shorten the testing time. Some items were deleted, and others were replaced by new items that were easier to administer.

Following analysis of the pilot test results, AIR recommended 156 early reading items and 79 mathematics items for inclusion in the full-scale 2004 field test.[17] The majority of test items recommended by AIR, both in the early reading domain and in mathematics, were selected from published instruments—*PreLAS 2000* (Duncan and De Avila, 1998), Peabody Picture Vocabulary Test–Third Edition (PPVT-III; Dunn and Dunn, 1997), and Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP; Lonigan et al, 2002) for the early reading test and Test of Early Mathematical Ability-3 (TEMA-3; Ginsburg and Baroody, 2003) for mathematics—and required copyright permission. The remaining recommended items came from the ECLS-K or were developed specifically for the ECLS-B.

AIR also provided general guidelines for testing preschool children, pacing the tests, and dealing with various child behaviors, and suggested language for assessors to use in explaining tasks to children and giving feedback.

2.2.2 Preschool 2004 Field Test

The proposed preschool child assessment was broad in its domain coverage, with items drawn from multiple standardized and/or widely used instruments (see table 8). The final selections for the 2004 ECLS-B field test instruments included 120 scored (i.e., nonpractice) items in early reading and 71 scored mathematics items (Burns et al. 2003). The item pools contained far too many questions for the total pool to be administered to any one child. For the purposes of the field test, items were divided within each domain into three forms: A, B, and C. Each child was administered one of the three forms. The forms were intended to be comparable in terms of coverage and difficulty. Therefore, items were distributed across forms such that each form contained a similar number of items in each domain, with each form including items of the same format at a similar difficulty level across forms. Altogether, each form contained 81 or 82 test items (there was some overlap among forms), with an additional 33 practice items, 1 or 2 introducing each of the various item formats.[18] However, not all children were administered all practice items; in most cases, if the child passed the first practice item in a set, the second one was skipped.

[17] A color knowledge test was proposed for inclusion in the preschool battery, but was not part of the pilot test.

[18] The ECLS-B assessment included items that required children to respond in different ways (e.g., verbal response, pointing to an image on the stimulus easel) to items of different formats (e.g., multiple choice, constructed response). Items requiring the same type of response to items of the same format were grouped together for ease of administration.

Table 8. Sources and numbers of items in early reading and mathematics domains, ECLS-B preschool field test: 2004–05

	Early reading items	Mathematics items
Total	120	71
PreLAS 2000	22	0
Peabody Picture Vocabulary Test–Third Edition (PPVT-III)	32	0
Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPP)	35	0
Elision subtest[1]	8	0
Initial Sound Matching Subtest	4	0
Test of Early Mathematical Ability-3 (TEMA-3)	0	2
Family and Child Experiences Survey (FACES)	7	1
Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K)	11	29
Original items	1	39

[1] The Elision subtest measures children's ability to remove phonological segments from spoken words to form other words.
NOTE: Items taken from existing instruments were administered as they were in the original instrument.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool field test, 2004–05.

All children received three subtests of the PreLAS: Simon Says, Art Show, and Let's Tell Stories. In addition to providing scoreable data for the language domain, these three PreLAS subtests were used to assess each child's English language proficiency. Only those children demonstrating English-language competency, as measured by their performance on these PreLAS subtests, were administered the rest of the early reading, color knowledge, and mathematics items in the direct child assessment.[19] In addition, each child who was not routed out of the cognitive assessment due to language was administered one of the six school readiness subtests of the Bracken Basic Concept Scale-Revised (colors, letters, numbers/counting, sizes, comparisons, or shapes) to validate the ECLS-B instrument. For children completing all components of the child assessment during the field test, total administration time was 45 minutes on average (15 minutes for the early reading section; 10 minutes for Let's Tell Stories; 17 minutes for the mathematics component; and 3 minutes for the Bracken subtest).[20]

The field test design called for 400 responses on each test item so that stable item parameter estimates could be computed and differential performance for at least some population

[19] All of the assessments in the field test were administered in English only.
[20] By design, all children were administered the PreLAS subtests first. Children demonstrating adequate proficiency to be assessed in English were then administered items from one of the three test forms. Test performance might be improved by a practice effect; that is, a child would perform better on items administered toward the end of a test because earlier items served as practice tasks. Conversely, if a fatigue effect is operating, children may do better on items administered near the beginning, before they have become tired. To compensate for the possible occurrence of these effects, the order of the early reading and math (including color knowledge) items was randomized so half of the children received the early reading items first, the Bracken subtest, then the math and color knowledge items; the other half received the math and color knowledge items first, then the Bracken subtest, then the early reading items.

subgroups could be assessed.[21] With three test forms, the field test required a sample of 1,200 children, each responding to test items in one of the three forms (A, B, or C). There were 650 field test participants with birthdays in January through March 2000 who had participated in the ECLS-B for field test purposes since about 9 months of age. A supplementary sample of 650 children was recruited specifically for the preschool field test of the cognitive assessment to ensure that at least 1,200 children would receive the assessment and at least 400 responses to the each item could be obtained. The supplemental sample was selected so that it included children born between April and December 2000, to ensure that the full field test sample would more closely represent the distribution of ages in the national sample. Of the 1,300 children recruited for the preschool field test, about 1,250 participated. Nearly all of the participants were at least 4 years old, with an average age of about 54 months (i.e., 4.5 years old).

2.2.3 Analyses of Preschool Field Test Data

Analyses of field test data focused on psychometric characteristics of the test items, the ability range of the field test sample, and operational issues. Psychometric analysis included calibration of item difficulty and discrimination, identification of flawed items that could be revised, and detection of items showing differential item functioning (DIF) with respect to population subgroups. Validation of the ECLS-B early reading and mathematics field test item pools was carried out by correlating field test ability estimates with subtests of the Bracken Basic Concept Scale. Although the field test sample was not designed to be nationally representative, the distribution and range of abilities found in the field test sample provided input in estimating the item difficulty range that would be required for the national sample assessments to ensure that the abilities of all children would be adequately measured. Operational issues examined included timing, completion rates, and feedback from field interviewers.

Three approaches were used in analyzing the psychometric characteristics of the field test item data: classical item analysis, item response theory (IRT), and DIF. The first two provide information on item difficulty and the relationship of individual items to the construct as a whole; the third analyzes differential performance by subgroup. Classical item analysis includes examining the percent correct (P+) for each item and the correlation of performance on each item to performance on the test as a whole (*r*-biserial). The *r*-biserials provide a convenient measure of the strength of the relationship of each item to the total construct.

IRT analysis can provide further information about the reliability of the items and their measurement properties by taking into account omitted items and the possibility of guessing. The IRT difficulty parameter "b" for each item is on the same scale as the ability estimate (theta) for each child, allowing for matching a set of test items to the range of ability of sampled children. The IRT discrimination parameter "a" is analogous to the *r*-biserial of classical item analysis.

[21] The ECLS-B assessment included items that required children to respond in different ways (e.g., verbal response, pointing to an image on the stimulus easel) to items of different formats (e.g., multiple choice, constructed response). Items requiring the same type of response to items of the same format were grouped together for ease of administration.

DIF analysis shows whether there is any relative advantage or disadvantage of a test item for children of different population subgroups who are matched on overall performance. In one method of DIF analysis, items are classified as "A" (no significant difference) to "C" (substantial and statistically significant difference for the focal group compared with a reference group). More detailed descriptions of these three techniques can be found in chapter 3 of this report.

The results of the field test analysis and their implications for the design of the preschool main studys are summarized below, first for the early reading assessment and then for mathematics. Details of the field test results may be found in the *Early Childhood Longitudinal Study, Birth Cohort: The Preschool Year (ECLS-B, Preschool Year) Field Test Report #5, Child Assessment* (Wallace and Dedek 2005).

2.2.3.1 Early Reading

As noted in section 2.2.2, the 120 scored early reading items included in the preschool field test were drawn from a number of sources, but were split into three test forms to reduce child response burden. All three forms had the same 22 PreLAS items; each form included 10 or 11 PPVT items (the number of PPVT items varied across the different field test forms) and 22 literacy items tapping letter and letter-sound knowledge, phonological awareness, print conventions, and word recognition. The PPVT and literacy items were different for each form.

The language items tended to be very easy for the field test sample. In the PreLAS set, 90 percent of children answered at least 8 out of 10 Simon Says items correctly and 96 percent of children answered at least 8 out of 10 Art Show items correctly. The 32 PPVT items field tested also were easy—about 88 percent of children correctly answered 8 or more of the PPVT items presented in their field test form. The national sample was expected to have similarly strong language skills since all of the children were sampled in the United States at birth. Thus, even those whose first language may not have been English would have had considerable exposure to English by the time of the preschool wave. The ceiling effect[22] and limited variance in the language measure captured by the PreLAS and PPVT meant that these items would serve primarily as a screening measure rather than a broad-range assessment of skills.

One goal of the field test was to better understand the administration and field scoring challenges associated with the PreLAS "Let's Tell Stories" items. The findings of the field test suggested several challenges in implementing these items and scoring them in the field. These items were problematic for several reasons:

Many children received low scores on Let's Tell Stories for reasons other than deficits in their English language fluency. During the initial field test period, about 100 children who came from homes where the parents reported that English was

[22] A ceiling effect is observed when data pool at the high end of the score range, so that rather than a normal or near-normal distribution, a severe right skew is seen in the data. A ceiling effect tends to suggest that an item set is too easy for those responding to the items. As noted in the text, more than 90 percent of children correctly responded to at least 8 of the 10 items in the Simon Says set and in the Art Show set.

spoken were classified as lacking English proficiency according to their overall score on the full set of PreLAS items (Art Show, Simon Says, and Let's Tell Stories). Poorer-than-expected performance resulted from a range of reasons. Some children who were highly articulate during other parts of the assessment or the home visit did not respond to the items in an articulate way. Some children refused to speak, cried, mumbled, or whispered to their mothers, while others spoke too quietly to be heard or put their hands over their mouths. Some felt stressed or threatened by the audiotape equipment, which then had to be hidden. These observations suggested that for these children at least, scores on this set of items were not accurate measures of the child's expressive English language skills. This was particularly problematic because the children exhibiting these behaviors were not of a particular subgroup but rather a random grouping which increased the variance associated with the PreLAS scores, thereby reducing the item reliabilities.

Field interviewers had trouble administering and recording the Let's Tell Stories properly. The interviewers audiotaped children's retelling of each of the two PreLAS stories, wrote down the child's version of the story, and then recorded a total score for each story on a 0–5 scale. Audiotapes captured instances of interviewers improperly prompting responses or modifying children's words, failing to pace children's responses in an appropriate manner, and failing to prevent interference or prompting by parents. The quality of the recordings was affected by background noise in the homes, equipment malfunction, and many children's unwillingness to speak clearly into the recording device.

Field interviewers had trouble reliably scoring the Let's Tell Stories in the field. Scoring reliability was unacceptably low as measured by relatively poor agreement of field interviewer ratings written down in the field with ratings by RTI staff that transcribed the audiotapes and scored the stories later. Discrepancies of as much as 3 points on the 0–5 scale were not unusual.

For all of these reasons, Let's Tell Stories did not fulfill the intended role as a valid and reliable measure of English fluency as administered in the field test.

Ability estimates (in the IRT theta metric) were obtained for the early reading field test to guide the selection of items with an appropriate range of difficulty levels required for the national test. The literacy items (letter recognition, print familiarity, etc.) included in the early reading assessment showed sufficient variability in difficulty to adequately assess the range of child ability levels expected in the national sample. The field test sample had a mean theta of -0.15 and standard deviation of 0.77 for the early reading sections. To the extent that the field test sample was representative of the national sample, test items with difficulty parameters two standard deviations below and above the mean theta, ranging from -1.69 to +1.39, were expected to provide accurate measurement for at least 95 percent of the national sample.

DIF analysis was carried out for the early reading field test, with a finding of C-level DIF for only one item (PPVT "trunk").[23] This item had been used in the ECLS-K assessment without significant DIF at that time. The DIF finding in the preschool field test was assumed to be related

[23] C-level DIF was found for this item comparing Black, non-Hispanic children with White, Non-Hispanic children.

to its small sample size, and the item was retained in the item pool. DIF analysis of the full national sample would be examined to determine whether the item would need to be deleted from scoring.

2.2.3.2 Mathematics

As was noted with the early reading field test instrument, the 71 scored mathematics items were split across three test forms, each containing 27 scored items. The results for the preschool field test of mathematics items highlighted issues that in some cases overlapped with those described above for the early reading test, as well as other issues unique to the mathematics section. In many instances, a practice mathematics item was followed by only one or two scored items before the task changed again. This resulted in inefficiency in the use of time as well as the potential for confusion. The three forms of the mathematics field test each required an average of 17 to 18 minutes to complete. Reducing the number of different presentation formats, as well as reducing the total number of items required to assess children, was expected to result in a preschool mathematics assessment that was, on average, well within the 15-minute time allowance.

Reviewers of the mathematics item pool cautioned that the length and complexity of the verbal load in the instructions and test questions could affect children's performance in the mathematics assessment. They suggested keeping the presentation of the items as simple as possible, using several practice items for some item types, and using very easy items for practice. By doing this, children's ability to understand what was expected of them (i.e., understanding the item format) would not be compromised by practice problems that were beyond their ability mathematically. Advisors also noted that some items (e.g., "What number did you just count?" or "What comes after eight when you count?") measured verbal memory more than mathematical concepts. A review of the mathematics field test analysis revealed that some of the constructs and presentation formats provided good measurement across a broad range of abilities, while others were appropriate only for the children in the sample with either the lowest or the highest ability. The field test results showed that the expectations in the mathematics framework for children's ability to manipulate numbers were too high. Some of the more difficult tasks were found to be unsuitable for all but the highest ability children.

The field test mathematics ability estimates (theta, in the IRT theta metric) averaged -0.14, with a standard deviation of 0.85.[24] Test items with difficulty parameters two standard deviations below and above the mean theta, ranging from -1.84 to 1.56, were expected to provide accurate measurement for at least 95 percent of the national sample.

[24] Note that each of the calibrations for reading and math were carried out independently. Thus, the mathematics metric has no relationship to that of the early reading section. These parameters have an interpretation only within the domain in which they were calibrated, not across domains.

Mathematics items were analyzed using DIF for all subgroup contrasts for which sample sizes allowed. No mathematics items showed significant or large differences between the subgroups when controlling for overall ability.

2.2.3.3 Item Validation

The preschool field test item pools were validated by concurrent administration of one of the six School Readiness subtests (i.e., colors, letters, numbers/counting, sizes, comparisons, and shapes) of the Bracken Basic Concept Scale-Revised. The Bracken is a receptive measure of these concepts that requires only a pointing response to indicate which pictured object is described (e.g., point to the largest one). Each child determined by the PreLAS to be sufficiently competent in English was administered one of the subtests. Two Bracken subtests were paired with each of the three test forms; one of the two subtests was randomly assigned to each child, resulting in approximately 150 to 200 observations on each subtest. The Bracken subtests were administered between the two primary subject matter assessments (early reading and mathematics). Correlations were calculated for each subtest score with the field test ability estimates for early reading and mathematics. The correlation of the Bracken Letters subtest with early reading was 0.82, while the correlation of the Bracken Numbers subtest mathematics field test ability estimate was 0.75. Correlations between other Bracken math-related subtests (sizes, comparisons, and shapes) were somewhat lower (ranging from 0.54 to 0.57), partly due to the composition of the field test item pool, which included many more items related to number concept than items related to size and shape (i.e., the two tests did not have similar content strand allocations). As a whole, recognizing that the different content strand allocations would result in somewhat lower correlations between the ECLS-B and Bracken items, these correlations suggest that the early reading and mathematics item pools displayed concurrent validity with accepted measures of early reading and mathematic skills.

2.2.3.4 Recommendations for the Preschool Assessment

After reviewing the field test results and consulting with experts, the following recommendations were made for the direct cognitive assessments:

Let's Tell Stories. It was recommended that this subtest not be used in the language screener, but be retained to provide an assessment of expressive language skills. To improve children's performance on the Let's Tell Stories subtest, it was recommended that the task be moved to after the early reading assessment in hopes that children would be more comfortable with the interviewer by then and, consequently, more talkative than they had been when the Let's Tell Stories was administered at the beginning of the cognitive assessment. Further, it was recommended that field interviewers not score Let's Tell Stories in the field. Rather, it was recommended that the audiotape of the assessment items and interviewer notes should be scored centrally by specially trained coders at RTI's secure research facility. Additionally, it was recommended that issues related to administration (e.g., interviewers improperly prompting responses, failing to properly prompt, problems

using the equipment, etc.) should be addressed directly in field interviewer training. Also, PreLAS coders at RTI should report any errors in administration that they encountered when coding the tapes and relay these errors to the field interviewer in a timely manner. These recommendations were implemented and, as a result, problems were minimized in the national preschool data collection wave.

Item Format. It was recommended that the main study assessment should include only the minimum number of format changes required to present domain content. Items in the preschool field test used many different presentation formats, including multiple formats to measure a single construct such as letter knowledge or phonological awareness. This resulted in inefficient measurement of skills because too much of the assessment time was used in retraining children to perform a task in a series of different ways and too little time was spent in actual performance and assessment of the children's skills. In selecting items for the main study, consideration was given to item format; where possible, items with consistent formats were selected over those with comparable content and difficulty but different formats.

Practice Items. It was recommended that the number of practice items be greatly reduced. In a number of sequences, two practice items were followed by only one scored item in the same format. This procedure not only was an inefficient use of assessment time but probably also resulted in some confusion for the child. Reducing the number of different item formats, which was implemented on the main study, reduced the need for some practice items. Also, where possible, practice items were sequenced so that if a child demonstrated adequate understanding of the item format in one practice item, no other practice items of that type were given.

Phonological Awareness Items. The literacy aspect of the early reading framework for the preschool assessment was somewhat optimistic with respect to expectations of preschool children's skills in the category of phonological awareness. These items as a group possessed weaker psychometric properties than other literacy items on the assessment. The framework specified that 18 of a total of 43 literacy items should relate to this construct. Results from the preschool field test showed that fewer than half the specified number of phonological awareness items would be satisfactory to adequately measure this construct. After discussion with experts, the number of phonological awareness items in the national preschool assessment was reduced to below the number specified in the framework.

Concept of Number. The framework stressed the importance of understanding that numbers represent objects; this concept was better measured by items requiring the child to count and relying on one-to-one correspondence than by the verbal memory items. That is, asking children to count stars and demonstrate that each star is assigned one and only one number (i.e., one-to-one correspondence) enabled them to better demonstrate their understanding of numeracy than those items that asked the child to say how many stars they had just counted (such items rely on verbal recall to succeed). Item statistics tended to corroborate this, showing weaker relationships with the overall numeracy construct for the verbal memory items than for counting items and mathematical concepts. Thus, the verbal memory items were dropped and more counting items were included in the assessment battery.

Manipulatives. Several of the formats used in the preschool mathematics field test asked the child to use plastic counters to solve a problem or extend a pattern. A series of tasks in which these objects were repeatedly taken out and put away interrupted the flow of the assessment. The design of the main study followed recommendations that items using counters be grouped together to limit disruptions during the assessment.

Test Structure. The original intent of the assessment design was to use an adaptive, two-stage test form where all children would receive a common set of items (a core or routing form) consisting of items across the ability distribution, and a second-stage form with items clustered to cover different portions of the ability distribution. The second stage allows for greater discrimination of children's ability than the core items, especially at low and high ends of the distribution. The analysis of the preschool field test data suggested that in the early reading content area, a single test form, with appropriate skip and/or discontinue rules, would be appropriate for assessing children's skills across the ability distribution. For the mathematics domain, however, a two-stage structure of a routing test followed by a second-stage form would be necessary to provide accurate measurement in the tails of the ability distribution, although for the majority of children a broadband single form could adequately assess their skills across most of the ability distribution. It was recommended that the original design of the assessment forms for the main study be modified to take this into account, and a single test form was developed for the early reading items, while a two-stage form was developed for the mathematics items (all children received a core set of items, with children responding to very few items correctly also receiving a basal set of items, and children responding to many core items correctly also receiving a ceiling set of items).

2.2.4 Preschool Wave Assessment

The preschool wave assessment consisted of (1) early reading; (2) mathematics; and (3) color knowledge. Color knowledge was not part of the field test work because the item used to assess children's color knowledge had been extensively fielded in other large-scale studies, such as the Family and Child Experiences Survey (FACES).

The assessments were untimed and individually administered. The preschool wave was also adaptive, in that the early reading assessment included a set of discontinue rules; after a child began encountering items that were too difficult, the section would be discontinued and the assessment would proceed to the practice items for the next construct, followed by scored items ordered from easy to hard. The mathematics assessment consisted of a core set of items, and, depending on a child's performance, he or she either ended with the core or was routed into a basal form (with an easier set of items) or a supplemental ceiling form (with a more difficult set of items).

Accurate measurement at all scale points requires that children receive sets of test items that are close to their ability level. The distributions of IRT ability estimates from the preschool field test sample were used to estimate the item difficulty for the main study and to define target difficulty ranges for different forms of the test (i.e., core, basal, and ceiling). Item quality and

potential for use were evaluated by reviewing all available information for each item. To contribute useful information about children's skill levels, test items selected for any assessment should ideally have high *r*-biserials (0.40 or higher) and IRT "a" (discrimination) parameters (at least 0.5, preferably 1.0 or higher), as well as good fits of empirical data to the IRT model. Items with high discrimination parameters permit accurate placement on the ability continuum because they identify (i.e., discriminate) finer differences in children's abilities than do items with low discrimination. For example, an item with a percent correct (P+) of 0.25 (a quarter of children answering correctly) and IRT "b" = 1.5 would appear to be a difficult item, potentially suitable for a high-difficulty test form. But a low *r*-biserial (below about 0.30 or 0.35) or a relatively flat IRT "a" parameter (below 0.50 or so) would suggest a weak relationship between the item and the test as a whole. In other words, although the item is difficult, it is not useful in differentiating different levels of skills.

The psychometric characteristics of the items were reviewed, and any items that were unsatisfactory with respect to the quality criteria described above were deleted.[25] For the remaining items, the difficulty statistics for the items within each content/presentation type were reviewed, and items were identified that would provide for an appropriate range of item difficulty across the battery.

Different presentations of the same content were compared, and where there was redundancy, the item sets with the strongest characteristics were selected. For example, items measuring the same construct, such as understanding number items, were presented in both receptive ("Point to the picture of three bananas.") and expressive ("How many bananas are there altogether?") styles in the field test. The preferred item presentation style was selected, for example, based on its IRT discrimination, *r*-biserial, and distractor analyses[26] (for multiple-choice items). The item presentation style exhibiting better results was prioritized for selection. In general, the types were ordered in increasing order of average difficulty (although most had a spread of difficulty within types), considering other factors such as grouping item types.

2.2.4.1 Early Reading

The early reading assessment consisted of a language portion and a literacy portion. The language portion comprised items that examine children's receptive language skills and vocabulary. A portion of the language assessment also was used to assess children's English-language proficiency. Based on their response to an initial set of 15 items administered in English, children were administered additional items in English or they were routed out of the English version of the assessment. A goal of the ECLS-B was to maximize participation of all

[25] A small number of the selected items fell short of these standards but were selected for other reasons, primarily framework specifications. In IRT, the measurement precision for individual examinees is improved by administering the maximum number of items possible in the time available and including items that function appropriately and measure the same construct.

[26] Multiple-choice items include the correct response, as well as a number of incorrect options (i.e., distractors). Ideally, distractors appear to be viable responses to children of low ability, but are not viable for children with high ability. Distractor analyses examine the frequencies with which children of differing abilities select the distractor(s) over the correct response. Items are considered to be flawed if a large proportion of children with high ability select a distractor on that item.

children in the English assessment; consequently, the threshold for passing the language items and being routed to the assessments in English was set low purposely. Depending on the parental indication of the child's knowledge of Spanish, the child was routed out of the child cognitive assessment completely (if not a Spanish speaker), or routed to a Spanish version of the assessment. There were about 100 children assessed in Spanish during the preschool wave, fewer than 50 children assessed in Spanish during the kindergarten 2006 wave, and no children assessed in Spanish during the kindergarten 2007 wave. [27] The Spanish assessments were built on Spanish- language versions of assessment items available from the test publisher, where available, and maintained the same structure as the English version.[28]

The literacy items examined children's phonological awareness, letter knowledge, awareness of the conventions of print, and word recognition. As stated, the literacy portion of the early reading assessment included a set of discontinue rules, whereby after a child began encountering items that were too difficult, the section would be discontinued and the assessment would proceed to the practice items for the next construct, followed by scored items ordered from easy to hard. Thus, there was just one early reading test with items spanning the full range of desired difficulties, as opposed to a routing test followed by a second-stage test. The number of items on the ECLS-B preschool early reading assessment that tap specific constructs is shown in table 9.

Table 9. ECLS-B preschool early reading assessment constructs, by number of items: 2005

	Number of items	(Number of practice items)
Total, language portion	15	(14)
Total, literacy portion	35	(7)
Phonological awareness	8	(4)
Letter sound knowledge and letter recognition	13	(2)
Print conventions	9	(0)
Word recognition	5	(1)

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

2.2.4.2 Mathematics

The mathematics items examined children's number sense, counting, operations, geometry, and pattern understanding. The mathematics assessment was a two-stage assessment, where all children were administered a core or common set of items, with supplemental items administered in a second stage only to children who performed very poorly on the core items (who received additional items at a much lower level) or very well on the core items (who received additional items that were more difficult than in the core).

[27] The Spanish assessments ultimately were not scored because only a small number of children took them, so IRT analyses necessary to develop scores were not psychometrically appropriate.

[28] Because some of the items on the child assessment were modified from standard instruments, or created specifically for the assessment, these items were translated into English, followed by back-translation.

The number of items on the ECLS-B preschool mathematics assessment that tap specific constructs is shown in table 10. The item types included in each form were the following:

Core: relative size/quantity, pattern matching, continue pattern of counters, counting, number recognition, ordinality.

Basal (low form): shapes, count fingers, count objects in pictures, count with counters.

Ceiling (high form): word problems with counters, word problems with pictures, and number sentences (abstract arithmetic).

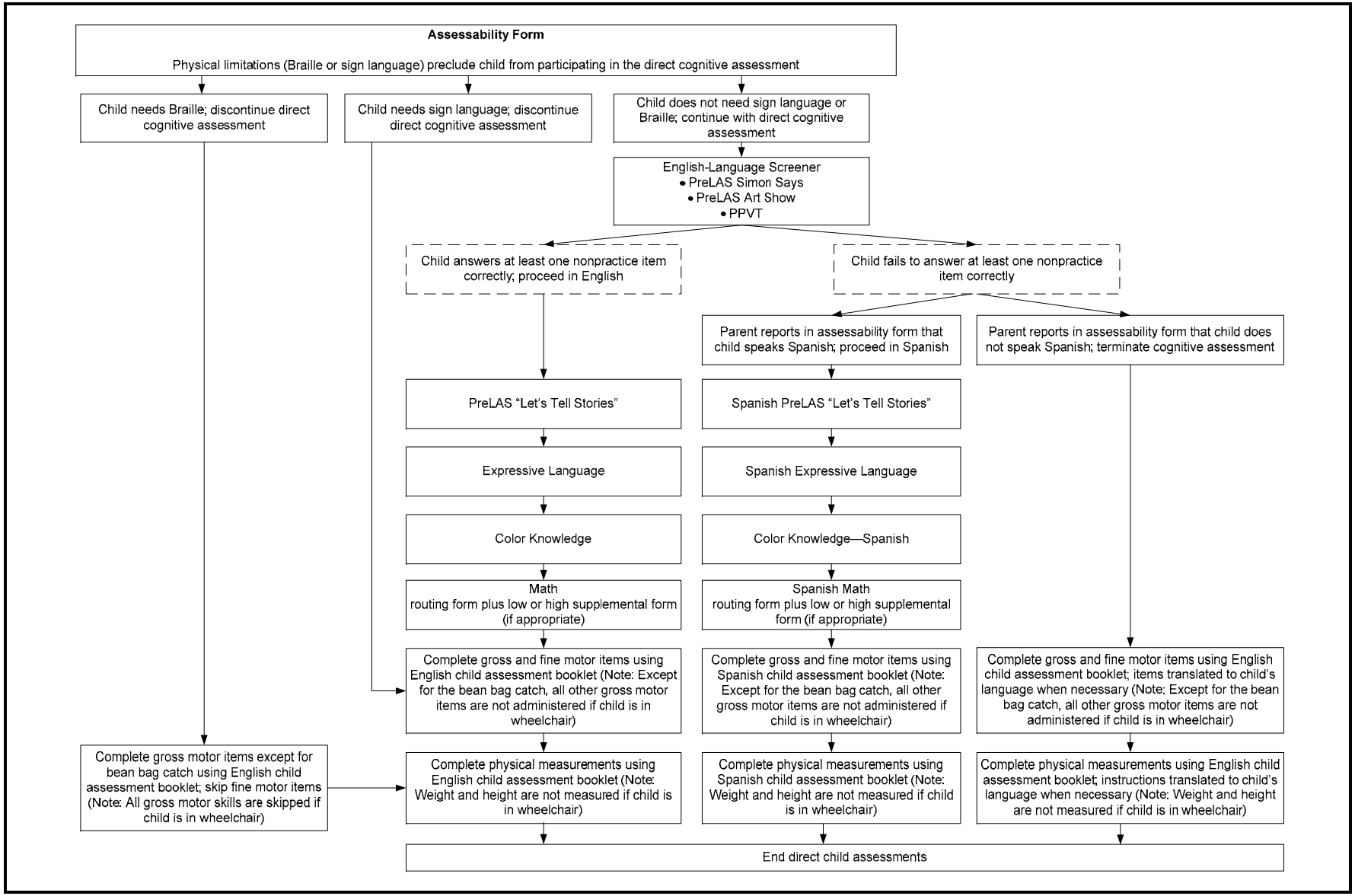
Table 10. ECLS-B preschool mathematics assessment constructs, by number of items: 2005

	Number of items	(Number of practice items)
Total	46	(15)
Number sense	10	(3)
Counting	14	(3)
Operations	8	(3)
Geometry	10	(4)
Pattern understanding	4	(2)

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

The flow of the preschool direct child assessment is shown in figure 2.

Figure 2. Flow of child assessment activities for the ECLS-B preschool data collection: 2005–06



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

2.2.4.3 Color Knowledge

The ECLS-B preschool assessment included items originally used as part of the Head Start Family and Child Experiences Survey (FACES, the "Color Bears" task) to assess a child's knowledge of colors. In this task, children were asked to name the colors of the 5 teddy bears (out of 10 pictured) indicated by the assessor. For all colors the child could not name in this way, the assessor provided the color name and asked the child to point to the bear of the color indicated. A child was given 2 points for items for which he or she could provide the color name and 1 point for items for which he or she could not provide the name but could correctly identify the color when the assessor provided the color name. Children who could do neither for a color received no points for that item. The color knowledge test was administered just prior to the mathematics assessment and is reported as a separate score.

2.3 Kindergarten Assessment (2006 and 2007)

The kindergarten wave assessment included early reading and mathematics (as previously mentioned; color knowledge was appropriate for the preschool wave, but not the kindergarten waves). This section presents an overview of the steps taken to develop the ECLS-B kindergarten assessment and information on item pool development, including a small-scale field test (section 2.3.1), analyses of field test data (section 2.3.2), and assembly of the main study (section 2.3.3).

2.3.1 Kindergarten Assessment Item Pool Development

The ECLS-B kindergarten item pool was a combination of items fielded as part of the preschool main study and items used with kindergartners in the ECLS-K.[29] To review item characteristics and functioning for the ECLS-B children when they were kindergarten age, a small field test with 303 children was conducted. A larger field test was not necessary because all of the items considered for inclusion in the assessment had been used before, either in the ECLS-B preschool wave or in the ECLS-K.

The final selections for the ECLS-B kindergarten[30] field test assessment instruments contained items in three early reading domains (language development, emergent literacy, and basic reading) and six mathematics domains (number sense, counting, operations, geometry, patterns, and measurement). The items included in the field test instruments were drawn from the ECLS-B preschool assessment and the ECLS-K kindergarten assessment. The early reading assessment contained 55 items and the mathematics test included 42 items. The assessment items were administered using guidelines provided by the original source for the item to ensure

[29] For detailed information on the ECLS-K assessment, see the ECLS-K Psychometric Report for Kindergarten Through First Grade (Rock and Pollack, 2002).

[30] Both kindergartners and first-graders were included in this field test to make sure that the items included in the ECLS-B kindergarten assessments would be challenging enough for use with two waves of kindergartners where the second wave might have somewhat older children than the first.

standardization. All of the approximately 300 children who participated in the field test completed both the early reading assessment and the mathematics assessment.

To create an item pool for the kindergarten assessment used in the 2006 and 2007 data collections, data from the ECLS-B kindergarten field test served as a supplement to data from the ECLS-B preschool items and the ECLS-K assessment items. By analyzing these data together, they can be put on a common scale allowing for measurement of longitudinal gains. To create the pool of items considered for inclusion on the ECLS-B kindergarten assessments, IRT calibration was carried out by pooling item-level data from the following datasets:

ECLS-B preschool wave (approximately 7,050 cases available at the time of analysis);

ECLS-B kindergarten field test (approximately 300 cases);

ECLS-K fall kindergarten (approximately 18,000 cases);

ECLS-K spring kindergarten (approximately 19,000 cases);

ECLS-K fall first grade (data collected for a subsample of about 5,000); and

ECLS-K spring first grade (approximately 16,000 cases).

Once the item pool was established by psychometrically establishing a link across the ECLS-B preschool assessment items, the ECLS-B kindergarten field test items, and the ECLS-K kindergarten and first-grade assessment items, the selection of items for the ECLS-B kindergarten assessment was informed by two sources of information: the difficulty parameters for each of the items in the item pool and the range of abilities expected among children in the kindergarten wave. Calibration of these two pieces of information *on the same scale*, so that they may be used in conjunction with each other, was accomplished by means of IRT analysis. (In-depth discussions of the application of IRT to longitudinal studies may be found in section 3.4.)

Overlapping items shared by two or more datasets served as an anchor for the common scale (i.e., the ECLS-B kindergarten field test helped to establish the performance of these overlapping items), such that parameters for items from different assessments and different samples could all be measured on a common scale. Pooling the datasets together also provided estimated values for the mean ability levels for each dataset on the same scale. Although the datasets were pooled, the samples were identified separately so the ability ranges of each dataset could be computed. Mean ability levels for each of the datasets listed above were calculated from the pooled sample. Using this information, an estimated ability range for children participating in the target administration, which in this case was for kindergartners, was determined.

2.3.2 Analyses of the Kindergarten Field Test Data

The ECLS-B kindergarten field test analysis focused on psychometric characteristics of the test items and the expected ability range of children in the kindergarten sample. As with the preschool field test, psychometric analysis used in the field test analysis included classical item

analysis, IRT, and DIF. A summary of these analytical techniques is provided with more detailed descriptions in chapter 3 of this report.

Fourteen early reading items were found to exhibit C-level DIF in the kindergarten field test analysis. Four Simon Says, three Art Show, six Peabody Picture Vocabulary Test (PPVT), and one print convention item exhibited DIF *against* various racial/ethnic subgroups compared with White, non-Hispanic children. Two mathematics items were found to exhibit C-level DIF on the basis of race/ethnicity when compared with the White, non-Hispanic reference group (one against Black, non-Hispanic children, and one against Asian groups). It was recommended that the 14 early reading and 2 math items demonstrating DIF based on race/ethnicity be excluded from the kindergarten design. Fortunately, other items in the same content areas (Simon Says, Art Show, PPVT, and print convention in early reading; measurement and spatial sense in mathematics) were available for inclusion in the kindergarten early reading assessment. Additionally, for all focal groups affected, there were other items that were relatively *easier* for the focal group compared with the reference group. These items, favoring the focal group, were not dropped from the item pool, as per standard psychometric procedures.[31] There was no evidence of DIF on any item based on child sex.

Finally, as noted in section 2.2.3.4, the assessment design was expected to utilize a two-stage adaptive design, with all children being administered a common set of items from across the ability distribution, with second-stage forms consisting of items clustered around different portions of the ability distribution administered to children (based upon their performance on the common items) to further define their ability levels. Analysis of item difficulty parameters from the kindergarten field test indicated that accurately measuring children's ability in both early reading and mathematics required the use of the two-stage design, and that items were available for inclusion in the assessment that were appropriate for the construction of both the routing form and second stage forms. Therefore, it was recommended that the structure of both the early reading assessment and the mathematics assessment be a routing test followed by a second-stage form. This structure allows the use of a routing form containing items drawn from multiple constructs within early reading and mathematics. The routing form is administered to all children to provide a common set of items across the difficulty range, while the second-stage forms are administered to children based on their performance on the routing form and contain items clustered at different areas of the difficulty range. Both the early reading and the mathematics assessments contained three second-stage forms of increasing difficulty (i.e., level 0, level 1, and level 2).

2.3.3 Assessment in the Kindergarten Waves Assessment (2006 and 2007)

The kindergarten assessment had a sufficient number of common items with the preschool assessment to establish a vertical link such that the preschool and kindergarten

[31] Note that this procedure differed from the DIF analyses performed on assessment data collected during the main study, where DIF analyses were conducted to evaluate potential bias for and against the focal group (see section 3.3).

assessment scores could be calibrated on the same metric. As with the preschool assessment, the kindergarten assessment began with language items, followed by literacy items. However, whereas the preschool early reading assessment had only one stage, the kindergarten early reading test was a two-stage test: a routing test followed by a second-stage test. The Let's Tell Stories task again was administered after the literacy items. The Let's Tell Stories items were modified from the preschool wave in that one of the two stories was replaced by a more difficult one (the other one remained the same). As planned, the color knowledge task was dropped. The assessment ended with the mathematics assessment.

2.3.3.1 Early Reading

The ECLS-B kindergarten early reading assessment was designed to be appropriate for use during both the kindergarten 2006 and kindergarten 2007 waves. The numbers of items in the ECLS-B kindergarten early reading assessment that relate to specific early reading constructs are shown in table 11.

Table 11. ECLS-B kindergarten early reading assessment constructs, by number of items: 2006–07 and 2007–08

	Number of items	Percent of items
Total	60	100
English language skills/oral language	7	12
Phonological awareness	15	25
Letter and letter-sound knowledge	14	23
Print conventions	6	10
Word recognition	11	18
Vocabulary	7	12

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

For the kindergarten early reading assessment, only those language items showing substantial variation in the preschool assessment were retained (resulting in fewer language items in the kindergarten assessment than there were in the preschool assessment). Additional items with higher difficulties, taken from the ECLS-K, increased the expected difficulty range of the kindergarten assessments, thereby allowing the assessments to accurately measure the abilities of the children in the two kindergarten waves (kindergarten 2006 and 2007).

The design of the assessment was driven by ability estimates derived from the ECLS-K fall kindergarten sample. Unlike the ECLS-K design, where all study children were in kindergarten during fall of the base year (the fall kindergarten wave), children in the ECLS-B entered kindergarten during fall 2006 or fall 2007. Children in the ECLS-B kindergarten 2006 sample was expected to be, on average, of lower ability than the ECLS-K sample, since the children in the kindergarten 2006 sample were on average younger than those in the ECLS-K, and about one quarter of the ECLS-B kindergarten 2006 sample was expected to have not yet

started kindergarten.[32] Children in the ECLS-B kindergarten 2007 sample were either in kindergarten for the first time or repeating kindergarten. As a result, the ability level of children in the ECLS-B kindergarten 2007 would be expected to be comparable to the ability level of children in the ECLS-K fall kindergarten sample, with some expectation of slightly greater ability for the ECLS-B children who were assessed later than fall of the kindergarten year (when compared with the ECLS-K fall kindergarten sample). Additionally, the average age of the ECLS-B kindergarten 2007 sample was expected to be slightly higher than the average age of the ECLS-K fall kindergarten sample.[33]

The ability (theta) estimates for the ECLS-K *fall kindergarten* sample, taken from the pooled analysis above, were used as a best estimate of the ECLS-B *fall kindergarten* 2006 sample, with a mean of -0.654 and standard deviation of 0.630. IRT ability estimates were used to define target difficulty ranges for different forms of the test. Therefore, it was important to extend the estimated ECLS-K fall kindergarten ability levels downward to avoid a floor effect for the kindergarten 2006 wave.[34] It was not necessary to extend the estimated ability levels upward to avoid a ceiling effect for the kindergarten 2006 wave, because the ECLS-B kindergarten 2006 sample was expected to be, on average, of lower ability level than the ECLS-K sample. The upper-bound from the ECLS-K estimate served to avoid a ceiling effect for the kindergarten 2006 wave. The ECLS-B kindergarten 2007 sample would be expected to have ability levels comparable to the ECLS-K sample or higher. Therefore, the estimated ability level would be extended upward to avoid a ceiling effect for the kindergarten 2007 wave. Thus, assuming that the ECLS-B average score would be equal to the ECLS-K mean, it would be expected that ECLS-B ability levels during the two kindergarten waves would fall three standard deviations above and below the ECLS-K mean (about -2.54 to +1.24), because this range represents an extension downward from the ECLS-K range while maintaining the same upper bound. This ability range was expected to include about 99 percent of the ECLS-B kindergarten 2006 group. For the lower end of the distribution, the estimated range also covered the ECLS-B preschool ability estimate to two standard deviations (0.568) below the preschool mean (-1.35), thus providing ample coverage for those children who would not be in kindergarten in fall 2006. Although the *average* ability levels of the ECLS-B group were expected to be lower than those

[32] The reason the ECLS-K kindergarten-year sample is older than the ECLS-B kindergarten sample is that the ECLS-K sample includes both first-time kindergartners and repeating kindergartners (who are typically older than most first-time kindergartners). The ECLS-B 2006 kindergarten-year sample includes only first-time kindergartners and children not yet old enough to enter kindergarten, or children who are age-eligible but not enrolled (i.e., "red-shirted").

[33] The ECLS-B kindergarten 2007 sample included children who entered kindergarten during fall 2006 (i.e., they were typically kindergarten age in 2006), but were repeating kindergarten in fall 2007 and children who were not old enough to enter kindergarten in fall 2006 because of birthdays after their localities' enrollment date. Both of these groups of children would tend to be slightly older than the average kindergartner.

[34] A floor effect is observed when item responses cluster at or near the very low end of the performance scale. This results in an apparent "floor" or lowest level of performance that may be due to the limited number of items for which children of lower ability may be able to respond correctly. Generally, when a floor effect is seen, the assumption is that the test is "too difficult" for many children, so items that are appropriate for lower levels of ability should be added.

from the ECLS-K,[35] designing the assessment to extend the range to three standard deviations above the ECLS-K average anticipated the possibility that there might be some high-ability ECLS-B children in the kindergarten waves (especially those in the kindergarten 2007 wave) who would require more difficult items for accurate measurement.

A range of -2.54 to +1.24 defined not only the expected ability range of the children but also the corresponding difficulty ("b") parameters of the items required for the assessment. Of course, the parameter estimates were not precise or permanent and were expected to change to some extent in the main study. There are many reasons for possible changes in the parameter estimates: the assortment of items and the order in which they would be given, the number and location of practice items, discontinue rules, and so forth. As a precaution against encountering floor and ceiling effects in the main study, the difficulty range for the items was extended at both the low and high ends. Some items with "b" parameters below -2.54 on the proposed low second-stage form and some above +1.24 on the high form were added.

The ECLS-B kindergarten wave early reading assessment began with a routing test followed by one of three second-stage tests (low, middle, or high). The psychometric characteristics of the items were reviewed, and any items that were unsatisfactory with respect to the quality criteria described above were deleted. For the remaining items, the difficulty statistics for the items within each content/presentation type were reviewed, and each item was classified as suitable for the routing test form, or a second-stage form—the low, middle, or high form—to provide for an appropriate spread of difficulty for each form. Different presentations of the same content were compared, and where there was redundancy, the item sets with the strongest characteristics were selected. In general, the types were ordered in increasing order of average difficulty (although most had a spread of difficulty within types), considering other factors such as grouping item types.

The reading portion of the assessment started with an English-language screener that was administered to all study children. These items were included to provide information on whether the child possessed sufficient English skills to understand the basic instructions and premises required to be assessed in English during the English reading and mathematics components, similar to what was done in the preschool wave assessment (see section 2.2.4.1).[36] Based on performance on this opening set of items, children either continued the assessment or were routed to a Spanish version (if they were Spanish-speaking) or other components of the assessment (i.e., physical measures and motor measures, which were administered to all children). The English screener items included five PreLAS Simon Says items (including two practice items), five PreLAS Art Show items (including two practice items), and five PPVT

[35] The ECLS-B kindergarten 2006 sample included children enrolled in kindergarten and those not yet enrolled, while the ECLS-K kindergarten year sample included only children enrolled in kindergarten. As a result, the expectation was that the ECLS-B sample would perform less well than did the ECLS-K sample.

[36] These language items were administered in English to all study children, regardless of the language spoken in the home, to provide a common metric for "competency" to be assessed.

items (including three practice items). As mentioned above, this set differed from those administered during the preschool wave. Items that showed little to no variation at preschool were not carried forward into the kindergarten wave assessment. To continue the assessment in English, the child had to correctly respond to at least one nonpractice item correctly.[37]

The distributions of thetas described above defined the range of abilities to be targeted by the test forms. The IRT difficulty parameters for the pool of available items were calibrated on the same scale as the abilities. Thus, the process of choosing test items relied on matching the difficulty of the items to the ability of the test takers. To optimize the measurement accuracy of the tests, the selected items needed to be approximately equally spaced along the ability/difficulty scale. Table 12 shows the expected peak difficulty ranges for the sets of items administered to the children; that is, the routing test plus one second-stage form (50 percent of the abilities were expected to fall in the routing + level 1 category, 25 percent lower [i.e., routing + level 0] and 25 percent higher [i.e., routing + level 2]); the number of items in the peak range; the full range of difficulty per form; and the total number of items per form. Again, to avoid floor and ceiling effects, the item difficulties in each of the forms needed to extend beyond the peak difficulty ranges. The items outside the peak range are a result of including the full range of routing items, the intentional addition of items to extend difficulties beyond the peak range, and the addition of items to provide the overlap between forms needed to support development of a common score scale.

Table 12. Peak and full difficulty ranges, routing + second stage, early reading: 2005–06

	Routing + level 0	Routing + level 1	Routing + level 2
Peak difficulty range	−2.54 to −1.07	−1.07 to −0.23	−0.23 to +1.24
Number of items in peak range	15 7 routing 8 level 0	22 6 routing 16 level 1	28 10 routing 18 level 2
Full range of difficulty	−3.04 to +1.25	−2.25 to +1.25	−2.25 to +1.64
Total number of items	40	45	51

NOTE: Item difficulty parameters are based on calibration of pooled field test data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten field test.

2.3.3.2 Mathematics

As noted in section 2.3.3.1, the ECLS-B assessment for the kindergarten waves was designed to be used during both the kindergarten 2006 wave and the kindergarten 2007 wave.

[37] A goal of the ECLS-B was to maximize participation of all children in the English assessment; consequently, the threshold for passing the language items and being routed to the assessments in English was purposely set low. As a result, very few children were not assessed in English. Children who did not answer at least one of these items correctly, but were reported by their parent to speak Spanish, received a Spanish version of the direct child cognitive assessment. Children who were routed out of the English cognitive assessment and who did not speak Spanish were routed into the noncognitive assessments. The Spanish assessments ultimately were not scored because only a small number of children took them, resulting in an inadequate sample size for conducting IRT analyses necessary to construct scores.

The numbers of items on the ECLS-B kindergarten mathematics assessment that relate to specific mathematics constructs are shown in table 13.

Table 13. ECLS-B preschool mathematics assessment constructs, by number of items: 2006–07 and 2007–08

	Actual number of items	Percent of items from actual forms
Total	58	100
Number Sense, Properties, and Operations	41	71
Measurement	3	5
Geometry and Spatial Sense	4	7
Data Analysis, Statistics, and Probability	3	5
Patterns, Algebra, and Functions	7	12

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections.

Similar to the early reading analysis, IRT ability estimates were used to define target difficulty ranges for different forms of the test. The ECLS-B sample during kindergarten 2006 was expected to be at or below the average ability level seen in the ECLS-K fall kindergarten sample, and the ECLS-B kindergarten 2007 sample was expected to have ability comparable to or slightly above that seen in the ECLS-K fall kindergarten sample.

The ability (theta) estimates for the ECLS-K *fall kindergarten* sample, taken from the pooled analysis, were used as best estimates of the ability of the ECLS-B sample when children entered *kindergarten* (i.e., the kindergarten 2006 and kindergarten 2007 waves), with a mean of 0.035 and standard deviation of 0.796. Once again, to the extent that the ECLS-K fall kindergarten test takers would be representative of the ECLS-B test takers, three standard deviations above and below the mean (about -2.35 to +2.42) was expected to include about 99 percent of the ECLS-B sample during the kindergarten 2006 and kindergarten 2007 waves. Note that the ability ranges of the children in the kindergarten 2006 sample were expected, on average, to be lower than those of the ECLS-K, so the ability estimates were extended at the lower end of the scale to include those children at the lowest ability levels. Children in the ECLS-B kindergarten 2007 sample would be expected to be at or slightly above the ability level of children in the ECLS-K fall kindergarten sample, so the ability range was extended (i.e., included up to 3 standard deviations above the ECLS-K average) to include those children at the highest ability levels. As a result, the mathematics ability range includes a lower bound below what was used in the ECLS-K, and an upper bound slightly above. This range also covered the (pooled) preschool ability estimate to 1.5 standard deviations (0.837) below its mean (-1.10), which was expected to provide ample coverage for those children in the lower ability range, such as those not entering kindergarten in fall 2006. Although the *average* ability levels of the ECLS-B group were expected to be lower than those from the ECLS-K, the assessments were designed in anticipation of the possibility that some high-ability ECLS-B children (especially those in the kindergarten 2007 wave) would need difficult items for accurate measurement. As

with the early reading forms, some items above and below the expected range were added to avoid floor and ceiling effects.

The item selection and form design processes for the mathematics assessment were similar to those of the early reading assessment. Items with unsatisfactory psychometric characteristics were removed. For the remaining items, the difficulty statistics for the items within each content/presentation type were reviewed, and each item was classified as suitable for the routing test form or a second-stage form—the low, middle, or high form—to provide for an appropriate spread of difficulty for each form. Different presentations of the same content were compared, and where there was redundancy, the item sets with the strongest characteristics were selected. In general, the types were sequenced in increasing order of average difficulty (although most had a spread of difficulty within types), considering other factors such as grouping item types. Test items for the kindergarten mathematics assessment were selected on the basis of the targeted range of abilities as described above. Table 14 shows the expected peak difficulty ranges for the sets of items to be administered to the children, the number of items in the peak range, the full range of difficulty per form, and the total number of items per form. To avoid floor and ceiling effects, the ability levels measured by each of the forms extend beyond the peak difficulty ranges.

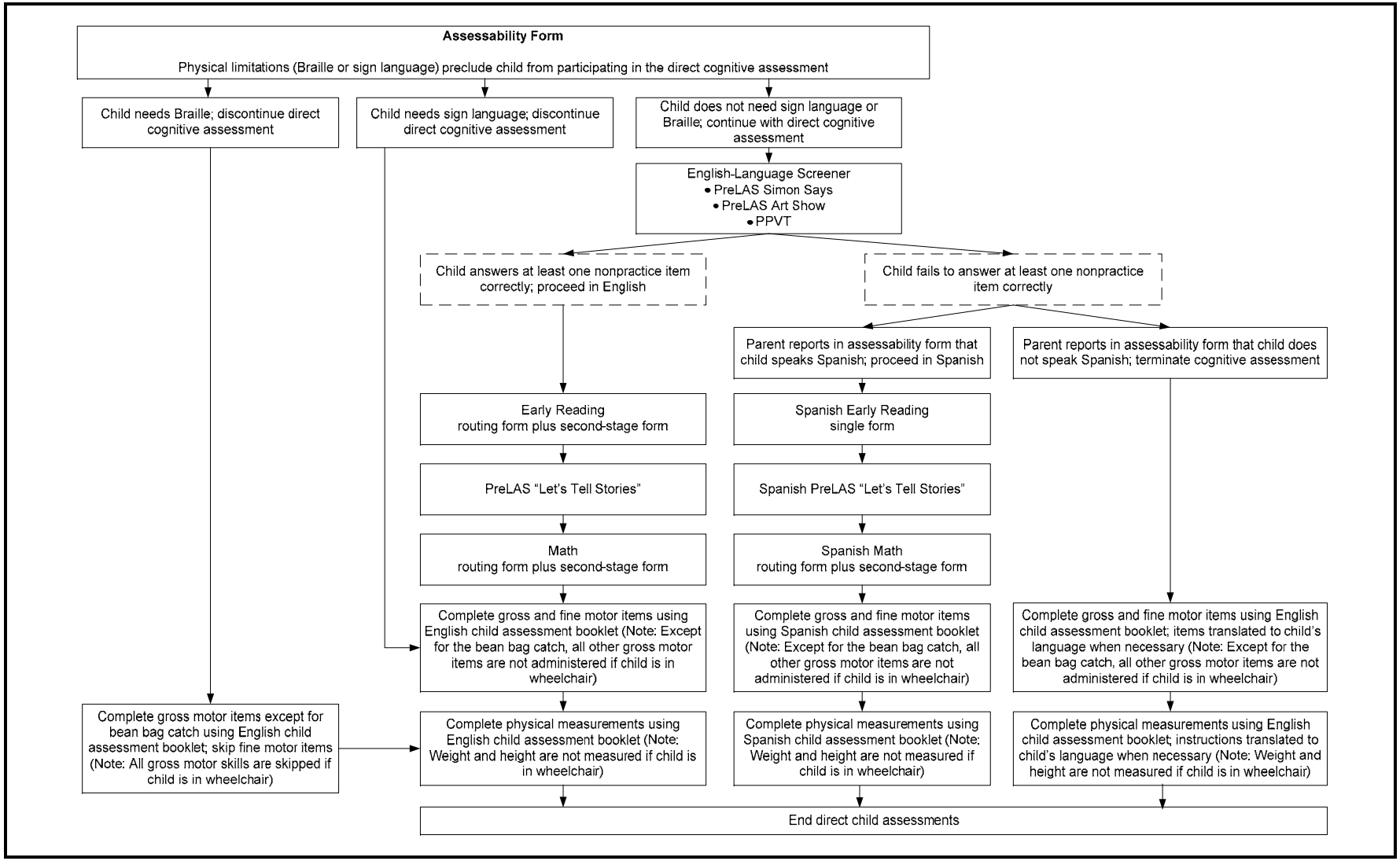
Table 14. Peak and full difficulty ranges, routing + second stage, mathematics: 2005–06

	Routing + level 0	Routing + level 1	Routing + level 2
Peak difficulty range	−2.35 to −0.50	−0.50 to +0.57	+0.57 to +2.42
Number of items in peak range	17 5 Routing 12 Level 0	16 6 Routing 10 Level 1	21 5 Routing 16 Level 2
Full range of difficulty	−2.83 to +1.77	−2.33 to +1.77	−2.20 to +2.83
Total number of items	33	37	42

NOTE: Item difficulty parameters are based on calibration of pooled field test data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten field test.

The flow of the direct child assessment during the kindergarten waves is shown in figure 3.

Figure 3. Flow of child assessment activities for the ECLS-B kindergarten 2006 and kindergarten 2007 data collection: 2006–07 and 2007–08



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections.

Chapter 3
Analysis Methodology

This chapter describes the procedures used to process data from the ECLS-B direct child cognitive assessments administered in the preschool and kindergarten waves and to produce scores for analysis. Quality control steps are described in section 3.1, followed by an explanation in section 3.2 of the methodology used to carry out specialized procedures for psychometric analysis. A three-parameter item response theory (IRT) model (Lord 1980) was used to put scores obtained on different sets of test items on the same scale for comparison within and across assessment years. Section 3.3 describes differential item functioning (DIF) procedures, which identified test items that performed differently for certain subgroups of the population and follow-up decisions made regarding the use of these items in the development of assessment scores included in the data file made available to researchers. Section 3.4 discusses the development of the longitudinal scales, including analysis of common items. The characteristics of the resulting scores for each wave are described in chapter 4.

3.1 Quality Control Procedures

Many procedures were employed to ensure accuracy in the collection of the cognitive test item data in the preschool, kindergarten 2006, and kindergarten 2007 data waves, including the review of timing data, item frequency reviews, and protocols programmed into the computer-assisted personal interview (CAPI) system designed to capture child responses to the assessment items. In the subsequent steps of converting the resulting raw item response data to final scores, procedures were used to ensure the accuracy and validity of the data. These steps included converting raw examinee item responses into scores for individual items, evaluating item functioning using both classical item analysis and IRT methods, and assembling item data into meaningful and interpretable scores. Throughout the process, attention was given both to checking that steps were carried out correctly and to verifying that results accurately represented the constructs they were designed to measure.

For each individual wave, frequency distributions of raw examinee item responses were produced for each test item to serve as a baseline for confirming the accuracy of later processing steps. Each distribution was compared with the text of the corresponding question in the assessment easel and with the instructions the assessor used in recording responses to confirm that responses were coded as expected. For example, for a four-option multiple-choice question, the data file would be expected to contain response codes of 1, 2, 3, and 4, while 1 (correct) or 2 (incorrect) was to have been recorded by the assessor for open-ended questions. Missing data codes (-7 = refused, -8 = "I don't know," -9 = nonresponse) also were counted for each item.

Before IRT analyses could be performed, a check was run within each domain (early reading or mathematics), so that children who had not responded to enough test items to receive

a score were identified and removed from analyses. "Too few items" was defined as answering fewer than 10 questions in the assessment for the domain.[38] When identifying unscoreable cases, codes for "I don't know" were treated as incorrect responses for open-ended items (those requiring a constructed response from the child without options to select), while conversely, codes for "I don't know" for multiple-choice items were *not* treated as valid responses because the child had the opportunity to guess and chose not to do so. Thus, only items actually attempted by the child were counted toward the scoreability threshold. Before being deleted from further analysis, each "too few items" data record was reviewed visually to verify that too few valid item responses were present. (The counts of children excluded for insufficient data are described in sections 4.2 and 4.3.)

Classical item analysis was carried out by individual wave for each test form using Educational Testing Service's (ETS's) proprietary software, F4STAT. Sets of statistics were produced for each item, as well as summary statistics for the form as a whole. Each of these statistics provided information on item performance and a source of quality control data. In terms of item performance, for each item, the number and percentage of test takers choosing each response option (or, for open-ended items, answering right or wrong) were computed, as well as the average number of correct answers on the whole test form for those test takers selecting a particular response option. The correct response for each item was identified in the statistical output file to facilitate analyses specifically involving the correct response, such as the one just described. Additionally, the same statistics were computed separately for children who omitted the item *and* answered at least one subsequent item ("omits") and for those who did not answer the item *or* any subsequent items ("not reached"). The response frequencies from the item analysis procedure were checked, item by item, against the baseline response frequencies initially obtained on the raw data file to confirm that responses and missing data codes had been interpreted correctly.

Summary statistics produced for each item included the proportion correct and *r*-biserial correlation of the item score (i.e., whether it was correct) with the total number-right score for its test section, adjusted to compensate for the attenuated correlation coefficient resulting from correlating a dichotomous variable (the item score) with a continuous variable (the total test score). These statistics were reviewed to verify that an unambiguous correct answer key was used for each item, meaning not only that the *intended* right answer was tagged, but also that the tagged answer was, in fact, functioning as an unambiguous right answer. Two indicators were used as evidence for the validity of the answer key: the mean section score for test takers choosing the correct response should be higher than that of the test takers choosing incorrect

[38] Because of the adaptive design of the assessments (not all children receive all items), and because some items administered as part of the early reading assessment were not included in the IRT calibration and scoring (see section 4.2.2), this check was performed to filter out those children with fewer than 10 responses on those items contributing to the IRT calibration. While children who answered fewer than 10 questions technically could be scored, when only a few items are available for a child, the likelihood of a stable estimate of child ability may become erratic, leading to problematic estimates and, possibly, unreliable estimates of the standard error of measurement. This is why the criterion of at least 10 answered questions was used.

responses, and the *r*-biserial should be positive, ideally at least .30 or higher. If these conditions are not satisfied, one of two error conditions could be responsible. An incorrect answer key could have been applied inadvertently, or the item may be flawed; that is, the intended correct answer may not really be correct, or there may be two or more equally correct response options. A low *r*-biserial also could be found for an item that is much too easy or much too hard for the vast majority of children. If virtually all children could answer an item correctly, or, at the opposite extreme, virtually all could only guess at the answer, the variance in item score would be low. Consequently, the resulting correlation with total test score (adjusted to compute *r*-biserial) also would be low. Because all of the items included in the preschool and kindergarten instruments had been field tested, and their response options had been evaluated and, if necessary, corrected, no flawed items were found.

Items within each test section or item type had been arranged in ascending order of anticipated difficulty. A review of an item's percent correct statistics would identify any serious deviation from this expectation, which could indicate anomalies in the administration or scoring of items. Similarly, unexpectedly large "omit" or "not reached" counts for an item or items could call into question whether routing steps or discontinue rules were applied correctly. No such indicators of data or administration errors were detected in reviewing item analysis tables for the kindergarten waves; one administration error in the preschool wave is described in section 4.3.1.1.

Summary statistics from the item analysis included the number of items and number of test takers analyzed for each section, the highest and lowest scores in each section, a measure of internal consistency (coefficient alpha reliability), and a frequency distribution of the number right for each section. Reliabilities were reviewed to confirm that they were consistent with expectations—typically about 0.80 or above for routing sections and for sections with more items, with lower reliabilities expected for second-stage forms (typically above 0.7), for which the restricted variance in overall ability (relative to the whole sample) would be expected to result in lower alpha coefficients, and for sections with relatively few items. The reliabilities for all test sections were consistent with these expectations. Sample and item counts and score ranges were checked for consistency with known values from administrative records (to confirm sample counts) and test forms (to confirm item counts and score ranges).

Because the mathematics assessment used during the preschool wave and the early reading and mathematics assessments used during the kindergarten waves had adaptive two-stage designs, an additional step was taken to examine data quality. Frequency distributions of routing test scores were compared with the distributions for each second-stage form to confirm that the cut points established during the assessment design phase had been implemented properly during data collection (i.e., that the number of observations for a particular second-stage form matched the number of observations with scores from the routing items in the score range that corresponded to that particular second-stage form). For example, in the kindergarten mathematics assessment, children who scored 5 or fewer correct on the routing part of the

assessment (not counting practice items) should have received the level 0 second-stage form, those with scores from 6 to 12 inclusive should have received the level 1 second-stage form, and those with 13 or more correct responses in the routing part of the assessment should have been routed to the level 2 second-stage form. Therefore, the number of children who scored 5 or fewer correct should have been the same as the number of children who were administered the level 0 second stage form, the number of children with scores from 6 to 12 inclusive should have been the same as the number of children who were administered the level 1 second-stage form, and the number of children with 13 or more correct responses in the routing part of the assessment should have been the same as the number of children who were routed to the level 2 second-stage form. Data records were reviewed visually to determine whether the counts reflected what was actually in the raw data files.

In addition to the classical item analysis results examined by separate wave, frequency distributions of total number correct (routing plus second stage combined) were examined separately for each form combination (routing + level 0, routing + level 1, and routing + level 2) to look for possible floor and ceiling effects. Although this is not a quality control issue in the sense of verifying the accuracy of the scoring procedures, it has implications for interpretation and analysis of the resulting scores. A floor effect occurs when the test is too difficult overall for many test takers and the score distribution contains a substantial number of children scoring at the chance, or guessing, level. Conversely, a ceiling effect occurs when a test is too easy for many children, and a substantial number of children are able to answer all, or nearly all, of the items correctly. Some floor and ceiling effects were found in the preschool wave data. This is discussed further in chapter 4.

The next step in processing the raw item responses was preparing scored item files for input to the IRT calibration procedures. These files were first prepared separately for each wave of collection. As part of this preparation, raw response option codes (e.g., 1, 2, 3, 4) were replaced with standard codes for “correct” (code = 1), “incorrect” (code = 0), “omitted” (code = 2), and “not reached” (code = 3) items. “Omitted” items were defined as unanswered items that were followed by a response to at least one subsequent item, whereas unanswered items were coded as “not reached” (or “not administered”) when the test had no subsequent items answered. In the early reading and mathematics assessments, the more difficult items at the end of a content section were not administered if a child had performed poorly on the easier items at the beginning of the section. Items that were omitted because of these specified skip patterns were coded as “not administered” rather than “omitted,” even though items later in the assessment, that is, in the next content category, may have been answered. The quality control procedure for confirming that this was done correctly consisted of printing the raw and scored data records for a spaced sample (i.e., equal intervals) of every 250th case, along with the answer keys, and hand checking for as many cases as necessary to confirm that the conversions were carried out correctly. In some cases, additional records were reviewed so that all variations found in the raw data file could be checked. For example, if the spaced sample of quality control

records happened to have only cases for children who were routed to the levels 0 and 1 second-stage forms, additional records were reviewed so that level 2 form score conversions could be verified as well.

Producing the scored item files entailed reorganizing the order of test items because some items appeared in more than one second-stage form. The scores for these common items needed to be relocated from their original separate locations to a single common location in the data file to subsequently permit linking during the IRT calibration. An item map was developed to direct the reordering of the common items. Once the items were reordered within the scored item files for each wave of collection separately, the kindergarten scored item files (from both waves) were combined with the scored item files from the preschool wave. Just as the test items shared in common across test forms within kindergarten had to be moved to a common location, items common across data collection waves were positioned together for IRT calibration and, again, frequency counts were checked to confirm the accuracy of the files. The non–IRT-based color knowledge score developed for the preschool wave was computed at this time, visually checked for accuracy in the same spaced sample, and inserted into the scored item files.

Finally, item-by-item frequency distributions were produced for the scored, reordered files; for the common items (i.e., those administered in more than one form within a wave or in more than one wave), the frequency counts were checked against the aggregates of the frequencies for the separate forms and waves in which the items originally appeared. These frequency counts, and item means computed on the verified score item files, provided the basis for checking the results of the IRT scaling steps.

Section 3.2.2 describes PARSCALE (Muraki and Bock 1991), the IRT program used for calibrating Bayesian estimates of item parameters and estimating test takers' ability levels on a scale that was then used to produce scale scores on the whole item pool. Statistics and graphs produced by the PARSCALE program and its associated graphing program (PARPLOT) were used not only to verify the accuracy of the computations, but also to evaluate the reasonableness of the results.

For each test item in the input scored data file, PARSCALE produced counts of the number of responses, number of omits, number right, number wrong, and percentage correct. These counts and percentages were checked, item by item, against the statistics generated from the scored, reordered data file to confirm that the correct input file was used and that the information it contained was read correctly by the PARSCALE program.

Another step taken for quality assurance, in addition to verifying the accuracy of the data and computations, was to evaluate the extent to which the scoring model appropriately represented the information in the whole item pool. The *r*-biserials produced in the classical item analysis steps showed the relationship of each test item with the rest of the form on which it appeared. Similarly, the IRT "a" parameter demonstrated the cohesiveness of the *whole set* of items used in each domain across the preschool and kindergarten assessments. High "a"

parameters (1.0 or above) were found for items strongly related to the underlying construct represented by the item pool. Approximately two-thirds of items in each domain had "a" parameters above 1.0. The more difficult items in each domain were administered to fewer children with similar ability levels, resulting in a low variability in item responses. Thus, discrimination between those who likely would or would not be able to answer an item correctly was not as finite with the smaller sample administered the more difficult items, resulting in "a" parameters below 1.0 for the more difficult items.

The graphs generated in conjunction with PARSCALE are a visual representation of the fit of the IRT model to the data. The modeled IRT parameters for each item define the shape and location of a logistic function for the item, which is plotted on a graph. Percentages of observed correct responses at intervals across the range of estimated ability levels are superimposed on the same graph. The closeness of fit of the data to the logistic function can be interpreted as confirming the appropriateness of the IRT model for scoring the tests. More detail on the IRT model is presented in section 3.2. Section 3.4 discusses the use and evaluation of the IRT procedures used in developing longitudinal scales across the preschool and kindergarten waves.

The final steps in producing the IRT-based scores consisted of aggregating probabilities of correct responses across the whole item pool in each domain for the scale scores at each wave. These scores were checked by printing a sample composed of every 1,000th case, including item and ability parameter estimates, and hand-checking computations. As a final check, means and standard deviations of the final scores were calculated and found to be consistent with the following expectations. For the scale scores, means were expected to increase from wave to wave, with a range of possible values that was consistent with the total number of items in the item pool for each subject (i.e., even though no child received all items, his or her predicted IRT scale score had the potential to indicate correct responses for all items).

3.2 Overview: The Three-Parameter Model

Measuring the extent of cognitive gains at both the group and individual levels required that the preschool and kindergarten assessment forms be calibrated on the same scale within each domain. IRT is the most efficient way to carry out such a calibration. IRT calibration requires that the sets of test items be relatively unidimensional within a domain with a single, continuous, trait (e.g., level of early reading ability) underlying all test form responses. To examine the sets of items for unidimensionality, factor analyses were run on the items selected for each assessment domain; a single, dominant factor was found for each domain.

There also should be a common set of anchor items shared by different forms or sets of questions, and most, but not necessarily all, content strands should be represented in forms from each wave. Sequential assessments (preschool, kindergarten 2006, and, as applicable, kindergarten 2007) must have increments in difficulty, which can be developed by (1) increasing the problem-solving demands within the same content areas across waves and (2) including

content in the later assessments that is more appropriate for children at a more advanced stage of development and builds on skills mastered earlier.

As indicated previously, IRT (Lord 1980) was used to calibrate the various forms within each content area. A brief introduction to IRT follows, with additional information on the Bayesian approach taken here.

3.2.1 Overview of Item Response Theory

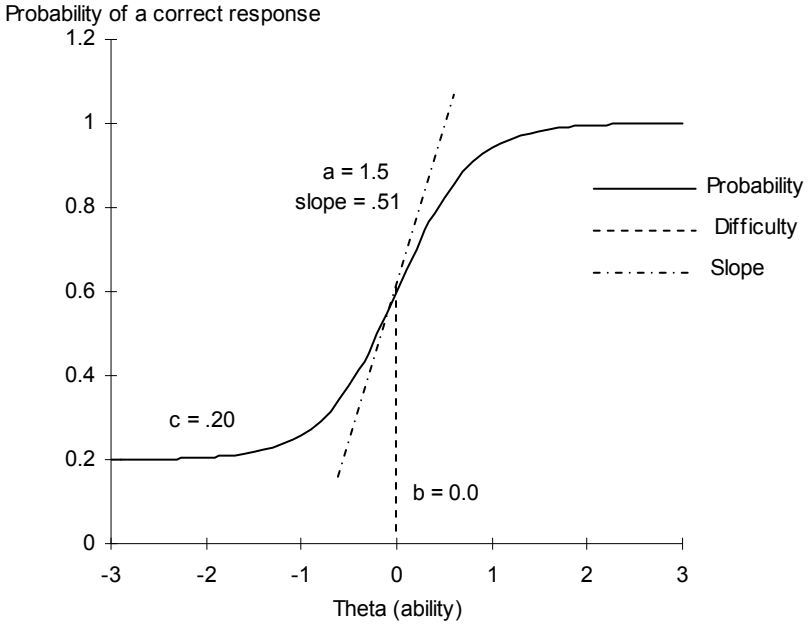
The underlying assumption of IRT is that a test taker's probability of answering an item correctly is a function of his or her ability level for the construct being measured and of one or more characteristics of the test item itself. The three-parameter IRT logistic model uses the pattern of "right," "wrong," and "omitted" responses to the items administered in a test form and the difficulty, discrimination power, and probability of guessing each item correctly, given the lowest level of ability, to place each test taker at a particular point, θ (theta), on a continuous ability scale. Figure 4 is an example of a graph of the logistic function for a hypothetical test item. The horizontal axis represents the ability scale, theta. Points along the vertical axis represent the probabilities of answering this item correctly given the level of ability (θ). The shape of the curve is given by the following equation, describing the probability of a correct answer on item *i,* or P_i, as

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-1.702 * a_i(\theta - b_i)}}, \tag{3.1}$$

where

θ = ability of the test taker;
a_i = discrimination of item *i*, or how well changes in ability level predict changes in the probability of answering the item correctly at a particular ability level;
b_i = difficulty of item *i*; and
c_i = "guessability" of item *i*, that is, the probability that a very-low-ability test taker will answer item *i* correctly.

Figure 4. Three-parameter IRT logistic function for a hypothetical test item

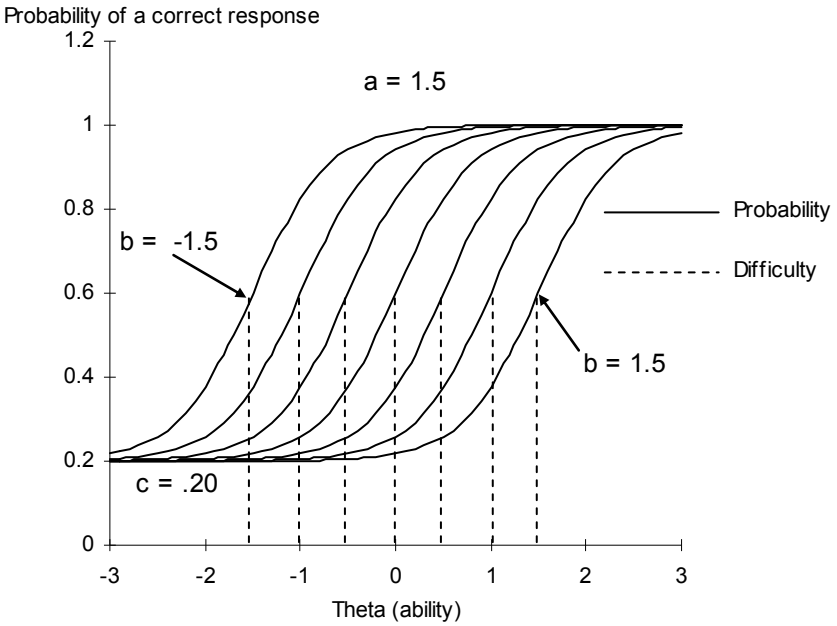


NOTE: a = parameter for discrimination; b = parameter for difficulty; and c = parameter for guessing. The discrimination parameter is proportional to the slope (tangent) of the function at the point of inflection.

The IRT "c" parameter represents the probability that a test taker with very low ability will answer the item correctly. In figure 4, about 20 percent of test takers with a very low level of mastery of the test material (i.e., a theta of -1 or less) guessed the correct answer to the question. Note that the "c" parameter does not necessarily equal 1/(number of options) (e.g., 0.25 for a four-choice item). Some response options may, for unknown reasons, be more attractive than random guessing, while others may be less likely to be chosen.

The IRT "b" parameters correspond to the difficulty of the items, represented by the horizontal axis in the ability metric. In figure 4, $b = 0.0$ means that test takers with $\theta = 0.0$ have a probability of getting the answer correct that is equal to halfway between the guessing parameter and 1. In this example, 60 percent of people at this ability level would be expected to answer the question correctly. The "b" parameter also corresponds to the point of inflection of the logistic function. This point occurs farther to the right for more difficult items and farther to the left for easier ones. Figure 5 is an example of a graph of the logistic functions for seven different test items, all with the same "a" and "c" parameters and with difficulties ranging from $b = -1.5$ to $b = 1.5$. For each of these hypothetical questions, 60 percent of test takers whose ability level matches the difficulty of the item are likely to answer correctly (i.e., 20 percent by guessing, plus half of the remaining 80 percent). Fewer than 60 percent will answer correctly at values of theta (ability) that are less than "b," and more than 60 percent will answer correctly at $\theta > b$.

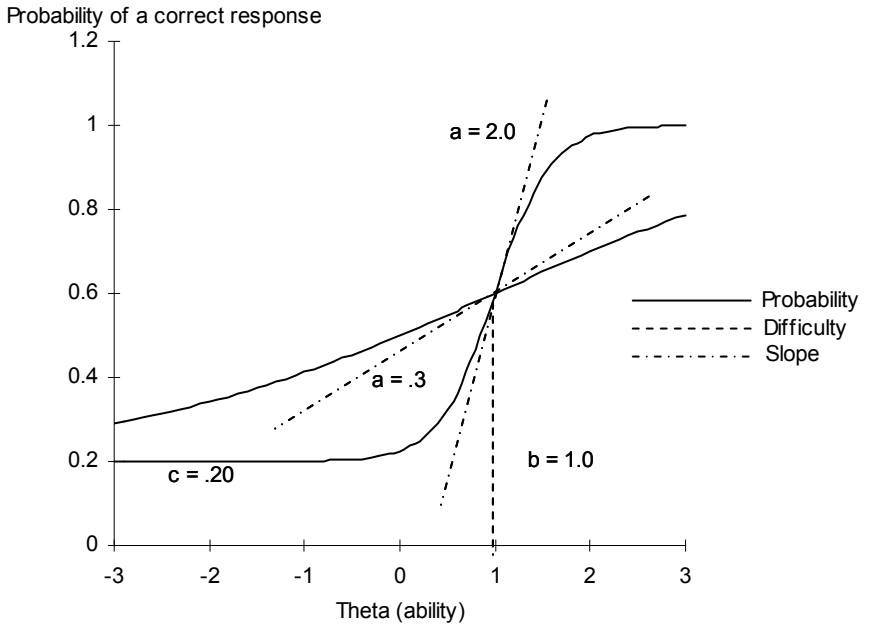
Figure 5. Three-parameter IRT logistic functions for seven hypothetical test items with different difficulty (b)



NOTE: a = parameter for discrimination; b = parameter for difficulty; and c = parameter for guessing. The discrimination parameter is proportional to the slope (tangent) of the function at the point of inflection.

The discrimination parameter, "a," has perhaps the least intuitive interpretation of the three IRT parameters. It is proportional to the slope of the logistic function at the point of inflection. Items with a very steep slope are said to discriminate well. In other words, they do a good job of discriminating, or separating, people whose ability level is below the calibrated difficulty of the item (who are much less likely to get it right) from those of ability higher than the item difficulty (i.e., "b"), who are much more likely to answer correctly. By contrast, an item with a relatively flat slope is of little use in determining whether a person's correct placement along the continuum of ability is above or below the difficulty of the item. This idea is illustrated by figure 6, representing the logistic functions for two test items having the same difficulty and guessing parameters but different discrimination. The test item with the steeper slope (a = 2.0) provides useful information with respect to whether a particular test taker's ability level is above or below the difficulty level, 1.0, of the item. A series of many such highly discriminating items, with a range of difficulty levels ("b" parameters) such as those shown in figure 5, will do a good job in narrowing the estimation of probable ability level. Conversely, the flatter curve in figure 6 represents a test item with a low discrimination parameter (a = 0.3). For this item, there is little difference in the proportion of correct answers for test takers who are several points apart on the range of ability. In this example, knowing whether a person's response to such an item is correct or not contributes relatively little to pinpointing his or her correct location on the horizontal ability axis.

Figure 6. Three-parameter IRT logistic functions for two hypothetical test items with different discrimination (a)



NOTE: a = parameter for discrimination; b = parameter for difficulty; and c = parameter for guessing. The discrimination parameter is proportional to the slope (tangent) of the function at the point of inflection.

With respect to evaluating item quality, "a" parameters should ideally each be over 0.50. Items with "a" parameters of 1.0 or above are considered very good. As described earlier, the "a" parameter indicates the usefulness of the item in discriminating between points on the ability scale. The "b" parameters, or item difficulties for the items, should span the range of abilities being measured by the test. Item difficulties should be concentrated in the range of abilities that contains most of the test takers. Test items provide the most information when their difficulty is close to the ability level of the examinees. Items that are too easy or too difficult for most of the test takers are of little use in discriminating among them. The "c" parameters (the probability of a low-ability person guessing correctly) tend to be about 0.25 or less for four-choice items, but they may vary with difficulty and, of course, the number of options. Open-ended items typically have a "c" parameter that is close to 0, or the "c" parameter may be constrained to be 0 if the probability of a correct random guess is very low. In general, the ECLS-B item parameters met these standards. Table 15 lists summary statistics for the "a" (discrimination), "b" (difficulty), and "c" (guessing) parameters for the item pools in early reading and mathematics.

Table 15. Early reading and mathematics item parameter statistics, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Score	*n*	Mean	SD	Low	High
Early reading					
"a"	85	1.61	0.95	0.28	4.91
"b"	85	0.73	0.99	-1.78	2.36
"c"	85	0.11	0.13	0.00	0.43
Mathematics					
"a"	71	1.32	0.61	0.34	3.24
"b"	71	0.22	1.25	-2.21	2.40
"c"	71	0.10	0.12	0.00	0.42

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections.

Once there is a pool of test items whose parameters have been calibrated on the same scale as the test takers' ability estimates, a person's probability of a correct answer for each item in the pool can be computed as a function of the person's ability estimate, theta, and the "a," "b," and "c" parameters for the item, even for items that were not administered to that individual. The IRT-estimated number correct for any subset of items is the *sum of the probabilities* of correct answers for those items. Consequently, the score is typically not a whole number.

In addition to providing a mechanism for estimating scores on items that were not administered to every individual, IRT has advantages over raw number-right scoring in the treatment of guessed and omitted items. By using the overall pattern of right and wrong responses to estimate ability, the IRT model gives little credit for correct answers to hard items by low-ability children, or to incorrect answers to easy items by high-ability children. Items considered "omitted" were presented to the child, with no response, and are treated as if the examinee had guessed at random. By contrast, raw number-right scoring, in effect, treats omitted items as if they had been answered incorrectly. Although the assumption of treating omits as incorrect may be reasonable in a motivated test for older children, it may not always be the case in the ECLS-B, where behavioral or other factors may contribute to a child's inability or unwillingness to complete all items. Therefore, IRT-based scores are preferable to number-right scores, and number rights scores are not developed.

3.2.2 Item Response Theory Estimation Using PARSCALE

The PARSCALE (Muraki and Bock 1991) computer program operates under the three main assumptions of IRT: (1) unidimensionality, (2) local independence, and (3) monotonicity. Unidimensional ability, also known as the latent trait, as measured by a test, can be used to make statements about the ability level of each test-taker. This is tested by factor analyses of the assessment items in each domain. The second assumption of local conditional independence requires that items are independent of each other given a particular ability level. This assumption is strongly related to the assumption of unidimensionality. Local conditional independence is often violated when the answer to a particular question depends on knowing the answer to

another question, especially when they are assessed in close, physical proximity. A clear example of local dependence is when a multiple choice question is followed up with a constructed response question asking the test taker to explain his or her answer. Such pairs of questions are, therefore, always scored as a single, combined question. Local independence was satisfied in the assessments. Finally, monotonicity is required (that is, the relationships between parameters are consistent). For the three-parameter IRT model described above, the probabilities of a correct response are defined by the logistic function and by the discrimination ("a"), difficulty ("b"), and guessing, or slope ("c") parameters. Item model fit to the observed data determines the validity of this assumption.

The PARSCALE program computes marginal maximum-likelihood estimates of IRT parameters that best fit the responses given by the test takers. The procedure estimates "a," "b," and "c" parameters for each test item, iterating until convergence when a specified level of accuracy is reached. Estimation-maximization steps are performed until the largest change in item threshold or slope parameters is less than the convergence criterion value, or the maximum number of cycles has been reached. The convergence criterion and maximum number of cycles are set using values defined by ETS standard practices (convergence criterion = 0.005 or 300 cycles). Comparison of the IRT-estimated probability of a correct response with the actual proportion of correct answers to a test item for examinees grouped by ability provides a means of evaluating the appropriateness of the model for the set of test data for which it is being used. A close match between the IRT-estimated probabilities and the empirical proportions indicates that the theoretical model accurately represents the empirical data.

A longitudinal growth study by its very nature consists of multiple subpopulations defined by abilities at differing times.[39] That is, after the preschool and kindergarten assessments had each been completed, there were three recognizable subpopulations of different ability levels related to time of testing (i.e., data collection wave). The preschool subpopulation will have, on average, a lower expected level of performance than that found for the same children a year (or two) later (during the kindergarten 2006 and kindergarten 2007 data collections).

At the time of the preschool assessments, very few children were able to answer some of the more challenging questions, such as those related to reading simple words or solving simple arithmetic problems. Because of skip rules in the preschool early reading test and routing specifications in mathematics, the most challenging items were not administered to children who had already failed to answer questions at an easier level. Thus, there were not enough data on the most difficult items to obtain stable item parameter estimates based on preschool data alone. (The IRT data sufficiency rule for the three-parameter model used in the ECLS-B calibration recommends a minimum of 1,000 observations per item for an assessment of approximately 60

[39] As used here, "subpopulation" refers to the data available at a point in time or around a given ability level (e.g., preschool or kindergarten). As used in IRT, subpopulation divides all available data across data waves (i.e., the "population") into smaller units based on differing levels of ability (i.e., "subpopulations"). In longitudinal studies, all children may contribute data into each subpopulation, because all children contribute data to the longitudinal data pool.

unique items.) However, many of these more difficult items were included in the kindergarten assessments and consequently readministered to some of the study children and administered to others for the first time. As a result, when the three data waves were combined for IRT analyses, the kindergarten 2006 and kindergarten 2007 data were used to stabilize the parameter estimates for the more difficult preschool items. Pooling data from the three time points and reestimating the item parameters essentially results in a remaking of history in a longitudinal study in which intermediate results are published before all the data are available. That is, preschool scores that had been analyzed and reported were later modified somewhat when the kindergarten data (from the kindergarten 2006 and 2007 data collections) became available. The use of all data points over time is desirable because it can provide updated, and more precise, estimates of both the item and latent child ability parameters throughout the entire ability distribution on a vertical scale. This procedure was used in the vertical scaling carried out for the National Education Longitudinal Study of 1988 (Owings 1995), for the High School and Beyond Longitudinal Study (Rock et al. 1985; Rock and Pollack 1987), and for the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) (Najarian et al. forthcoming; Pollack et al. 2005; Rock and Pollack 2002).

A strength of the PARSCALE and other Bayesian approaches to IRT is that they can incorporate prior information (i.e., data from previous data waves) about the ability distribution in the current wave ability estimates. This is particularly crucial for measuring change in longitudinal studies. It provides an acceptable way of coping with perfect scores (i.e., correct answers to all items administered) and chance scores (i.e., scores at the guessing level or below). For example, a few advanced individuals who take the preschool mathematics form might get all of the items correct. These children, while gifted, may *not* get perfect scores when they eventually are tested on a harder set of items in later waves. Conversely, individuals scoring at or below the chance level at two time periods will exhibit some gains but may have also gained skills that were below the level assessed by the original test items, and are therefore not measurable. Pooling all available information—that is, pooling all item responses for all children at all time points, and recalibrating all of the item parameters using Bayesian priors[40] reflecting the ability distributions associated with each particular wave—provides for an empirically based adjustment of item parameters and ability scores to values more representative of the population than the data from one wave taken in isolation might suggest (Muraki and Bock 1991). Bayesian priors (also typically referred to as simply priors) are essentially a priori distributional assumptions about proficiency and have relatively little influence on the proficiency estimation if there is sufficient information collected from a child, but has more influence if the child's information is sparse.

Using the total item pool in conjunction with the Bayesian priors (which reflect the ability distributions associated with each wave) leads to a reduction in extreme values of item

[40] A prior as used here is a proficiency (i.e., ability) distribution defined a priori to reflect prior beliefs of the true distribution. In this case, the proficiency distribution is believed to be standard normal, thus the prior is a standard normal distribution.

parameters, as well as an adjustment of the perfect and chance scores. This, in turn, allows for the potential for some gains even in the upper and lower tails of the distribution. Each wave of data collection (preschool, kindergarten 2006, and kindergarten 2007) is treated as a separate subpopulation with its own ability distribution. The amount of shrinkage toward the mean is a function of the distance from the subpopulation means and the relative reliability of the score being estimated (i.e., ability estimates in the tails of the distribution move more toward the mean than do those that are near the mean). For example, if the dispersion of the ability estimate is greater in one wave compared with another, the extremes of the ability estimate in the wave with the wider distribution will be shrunk more, in an effort to create more realistic estimates. Theoretically, this approach has much to recommend it. In practice, the model has to have reasonable estimates (i.e., better estimation of outliers in the ability distributions) of the difference in ability levels among the subpopulations (different data collection waves) to incorporate realistic Bayesian priors for the ability and item parameter estimates. The PARSCALE program generates initial item parameter estimates from default values or item difficulty statistics of a Bayesian prior calculation with a similar, or the same, population. Similarly, item parameter Bayesian priors and a priori distributions of abilities by subpopulation may be generated by PARSCALE or input from Bayesian prior distributions. Essentially, the longitudinal scales are determined by the linking items (i.e., items that are presented in more than one data wave and hence serve to link the data on a common metric), and the initial Bayesian prior ability means for the children in the different waves are in turn determined by the differential performance of the children on these linking items across waves. For this reason, the item pool has been designed to have an overabundance of items linking the forms across waves. This approach, using adaptive testing procedures combined with Bayesian procedures that allow for the use of prior values on both ability distributions and on the item parameters, is needed in longitudinal studies to minimize ceiling and floor effects.

A multiple group version of the PARSCALE computer program that was developed for the National Assessment of Educational Progress (NAEP) allows for both group ability priors[41] and item priors. A publicly available multiple group version of the BILOG (Mislevy and Bock 1982) computer program called BIMAIN (Muraki and Bock 1987, 1991) has many of the same capabilities for dichotomously scored items only. Because the PARSCALE program was applied to dichotomously scored items in the ECLS-B vertical scaling, its estimation procedure is identical to the multiple group version of BILOG or BIMAIN. PARSCALE uses a marginal maximum likelihood estimation approach and thus does not estimate the individual ability scores when estimating the item parameters but assumes that the ability distribution is known for each subgroup. Thus, the posterior distribution of item parameters is proportional to the product of the likelihood of observing the item response vector, based on the data and conditional on the item parameters and subgroup membership, and the assumed prior ability distribution for that

[41] There is a difference between population and item priors. The first set is across the whole population and is not related to the items.

subgroup. More formally, the general model in terms of item-parameter estimation is the same as that used in NAEP and described in some detail by Yamamoto and Mazzeo (1992, p. 158) as follows:

$$\begin{aligned} L(\beta) &= \prod_g \prod_{j:g} \int_\theta P(x_{j:g} \mid \theta, \beta) f_g(\theta) d(\theta) \\ &\approx \prod_g \prod_{j:g} \Sigma_k P(x_{j:g} \mid \theta = X_k, \beta) A_g(X_k). \end{aligned} \tag{3.2}$$

In equation (3.2), $L(\beta)$ is the marginalized likelihood of observing a given response matrix (students by items); $P(x_{j:g} \mid \theta, \beta)$ is the conditional probability of observing a response vector $x_{j:g}$ of person j from group *g*, given proficiency θ and vector of item parameters $\beta = (a_1, b_1, c_1, ..., a_k, b_k, c_k)$, for *k* items, each with discrimination parameter *a*, difficulty parameter *b,* and guessing parameter *c;* $f_g(\theta)$ is a population density for θ in group *g*; and $d(\theta)$ is the variable of integration. Prior distributions on item parameters can be specified and used to obtain Bayes modal estimates of these parameters (Mislevy and Bock 1982). The proficiency distribution can be assumed known and held fixed during item parameter estimation or can be estimated concurrently with item parameters.

The $f_g(\theta)$ in equation (3.2) are approximated by multinomial distributions over a finite number of quadrature points, where X_k for $k = 1, ..., q$, denotes the set of points, and $A_g(X_k)$ are the multinomial probabilities at the corresponding points that approximate $f_g(\theta)$ at $\theta = X_k$. If the data are from a single population with an assumed normal distribution, Gauss-Hermite quadrature procedures provide an optimal set of points and weights to best approximate the integral in equation (3.2) for a broad class of smooth functions. For more general population density function f or for data from multiple populations with known densities, other sets of points (e.g., equally spaced points) can be substituted, and the values of $A_g(X_k)$ may be chosen to be the normalized density at point X_k (i.e., $A_g(X_k) = f_g(X_k) / \Sigma_k f_g(X_k)$).

Maximization of $L(\beta)$ is carried out by an application of an expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). When population densities are assumed to be known and held constant during estimation, the algorithm proceeds as follows. In the E step, provisional estimates of item parameters and the assumed multinomial probabilities are used to estimate expected sample sizes at each quadrature point for each group (denoted $\hat{N}_{gk}$), as well as over all groups (denoted $\hat{N}_k = \Sigma_g \hat{N}_{gk}$). These same provisional estimates are also used to estimate an expected frequency of correct responses at each quadrature point for each group (denoted $\hat{r}_{gik}$) and over all groups (denoted $\hat{r}_{ik} = \Sigma_g \hat{r}_{gik}$). In the M step, improved estimates of the item parameters, β, are obtained using maximum likelihood by treating the $\hat{N}_{gk}$ and $\hat{r}_{ik}$ as known, subject to any constraints associated with prior distributions specified for β.

The user of the multiple group version of PARSCALE has the option of fixing the priors on the ability distribution or allowing the posterior estimate to update the previous prior and combine with the data-based likelihood to arrive at a new set of posterior estimates after each major EM cycle. If one wishes to update on each cycle, one can continue to constrain the priors to be normal or allow their shape to vary. The ECLS-B approach was to allow for updating the prior but with the normality assumption. The smoothing that came from the updated normal priors led to ability distributions that looked less jagged. If the updated ability distribution were allowed to take any shape, rather than being constrained to a normal distribution, lack of fit in the item parameter distribution would simply be absorbed in the shape of the ability distribution. A similar procedure was used in estimating the item parameters in the National Adult Literacy Study (Kirsch et al. 1993).

The solution to equation (3.2) finds those item parameters that maximize the likelihood across three points (preschool, kindergarten 2006, and kindergarten 2007). The present version of the multiple group PARSCALE saves the subpopulation means and standard deviations and the individual expected a posteriori (EAP) scores. The individual EAP scores, which are the means of the posterior distributions of theta[42] were obtained using the Gaussian quadrature procedure. This procedure is virtually equivalent to conditioning (e.g., see Mislevy, Johnson, and Muraki 1992) on a set of "dummy" variables defining the ability subpopulation from which an observation comes. The one difference is that the group variances are not restricted to be equal as in the standard conditioning procedure.

Conditional independence, or the assumption that any two items are unrelated to each other, conditional on proficiency, is an assumption of most commonly used IRT models, but as Mislevy, Johnson, and Muraki (1992) point out, it is a strong assumption that is often violated in practice. To ensure that there were no substantive violations of this assumption, factor analyses were carried out on the preschool and kindergarten item response data in early reading and mathematics to confirm that there was a large dominant factor underlying each content area. Finding additional small factors that are representative of only a subset of items could indicate that there is some dependence between items beyond the dominant factor and, hence, the local dependence assumption would have been violated. In addition, all graphs were inspected to ensure a good fit throughout the ability range. For each test item, the empirical proportion correct in each wave was computed and compared with the model-based estimated proportion correct based on thetas for the same set of children, that is, the subset of children in the wave who had

[42] The theta reported on the data file for each child is the mean of the posterior distribution of theta for that child. This single value and its associated SEM are reported for all eligible children on the data file.

received and responded to the item.[43] Discrepancies between predicted and actual item proportion correct were reviewed for each wave, and no systematic over- or underprediction was found for any type of item.

3.2.3 Standard Errors of Measurement Using the Information Function

In statistics and psychometrics, the precision of parameter estimates can be measured using the information function. This is computed as a function of the reciprocal of the measurement error, or the variability of repeated estimates of the value of the parameter, denoted as σ^2. Thus, the less measurement error is present, the more precise the estimate of the value a parameter, and the greater the value of the information function. Equation 3.3 defines the information function (I):

$$I = \frac{1}{\sigma^2} \tag{3.3}$$

In IRT, of interest is estimating the ability parameter, or θ, of each child. If the test contains a large number of highly discriminating items whose difficulty is appropriate for a particular child, the child's true ability can be measured with great precision. Measurement error will be low, and the value of the information function will be high. Conversely, if most of the test items are too difficult for a low-ability child, or too easy for a high-ability child, a precise estimate of the child's θ, ability level, cannot be obtained. The variance of estimates (measurement error) will be relatively high, and the value of the information function relatively low. Therefore, the information function tells how well each child's ability is being estimated.

Under IRT theory, each item on the test contributes to measurement of the underlying trait. Highly discriminating items (i.e., items with high "a" parameters) that are of appropriate difficulty for an individual child are most useful in pinpointing a child's ability level; items that are much too easy or much too hard, or that have low discrimination parameters, contribute relatively little. An item information function is computed for each item answered by a test taker. Since the overall test is used to estimate the ability level of the child, the test information function (sum of the item information functions) is used to estimate the standard error of measurement. The test information function is defined by:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{3.4}$$

[43] The mathematics item responses satisfactorily fit a single underlying factor. However, evidence from factor analyses at preschool, in conjunction with results of initial IRT calibration, pointed to the necessity of splitting the early reading test into separately scored sections (one for language, one for literacy) for the preschool wave. When kindergarten 2006 data were coupled with preschool data it was possible to combine the literacy and PPVT items (and not those from the PreLAS Art Show and Simon Says item sets) to produce an IRT calibration model that empirically and conceptually fit the data from both waves well, in addition to providing the means to develop a score similar to that of the ECLS-K reading measure. For more details, see chapter 4 of this report.

where

$I(\theta)$ = amount of test information at child's ability level θ;

$I_i(\theta)$ = amount of test information at child's ability level θ for item i; and

n = number of items answered by the child.

The test information function will be much greater than any single item information function, thus, a test measures ability more precisely than does a single item. The test information function is calculated using only the administered items with valid responses. Generally, the more items *answered* then, the greater the precision in estimating the ability.

The definition of the item information function depends on the IRT model used. For the three-parameter (a, b, and c) model used in the ECLS-B estimates and described above, the item information function is defined as:

$$I_i(\theta) = a^2 \frac{Q_i(\theta)(P_i(\theta) - c)^2}{P_i(\theta)(1-c)^2} \tag{3.5}$$

where

$P_i(\theta) = c + (1 - c)\frac{1}{1 + e^{-L}}$;

$L = a(\theta - b)$; and

$Q_i(\theta) = 1.0 - P_i(\theta)$.

The test information function is defined as the sum of the item information functions for each administered item at the child's given ability level. Tests are designed with item difficulties that are matched to the expected ability levels of the target population of test takers. There are generally more middle-difficulty items, matching the ability of the majority of test takers, and relatively few easy and difficult items designed for the children in the tails of the ability distribution. As a result, the abilities in the center of the scale are estimated with more precision than those in the tails.

The standard error of estimation is computed from the reciprocal of the square root of the test information function:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{3.6}$$

The procedure above was carried out to calculate a standard error of measurement for each of the theta estimates. These standard errors of measurement are reported in the data files for each of the thetas in early reading and mathematics for preschool, kindergarten 2006, and kindergarten 2007. The results of the standard error calculations for all waves are presented in chapter 4.

3.3 Differential Item Functioning

As defined here, differential item functioning (DIF) attempts to identify those items showing an unexpectedly large difference in item performance between a focal group (e.g., Black children) and a reference group (e.g., White children) *when the two groups are "blocked," or matched on their total score*. Any such strictly internal analysis (i.e., without an external criterion) cannot detect bias when that bias pervades all items in the test (Cole and Moss 1989). It can only detect differences in the relationships among items that are anomalous in some group in relation to other items. In addition, such approaches can only identify the items for which there is unexpected differential performance; they cannot directly imply bias. A determination of bias implies not only that differential performance on the item is related to subgroup membership, but also that the difference is *unfairly* associated with subgroup membership. That is, the difference is because of an attribute not related to the construct being measured. As Cole and Moss (1989) point out, items so identified must still be interpreted in light of the intended meaning of the test scores before any conclusion of bias can be drawn. It is not entirely clear how the term "item bias" applies to academic achievement measures given to children with different patterns of exposure to content areas. For example, some children may be in preschools where the curriculum emphasizes learning letter names and sounds, while others are in schools, or with day care providers or families, where relatively more time is spent reading stories to the children than teaching academic skills. Both groups may have similar total scores in early reading, but for one group the letter recognition items may be differentially difficult, while the reverse is true for the other group. In general, it is ETS's practice to carry out DIF analysis on tests it designs to detect test items with differential performance for subgroups defined by sex and race/ethnicity. In the ECLS-B, DIF analyses were also conducted on kindergarten enrollment status to determine any bias in items related to children's direct experience of schooling.

Two DIF methods were used in detecting differential performance of subgroups on the ECLS-B direct cognitive assessments during the preschool and kindergarten waves. One method is based on the Mantel-Haenszel (M-H) odds ratio (Mantel and Haenszel 1959) and its associated Chi-square. The other method uses a proportion correct difference metric and is commonly referred to as the standardized primary item discrepancy index (P-DIF) (Dorans and Kulick 2006). The two methods complement one another in detecting differential performance. The methods and advantages of using both procedures are discussed in the following paragraphs.

The M-H DIF program developed at ETS (Holland and Thayer 1986) forms odds ratios from two-way frequency tables. For example, in a 20-item test, 21 two-way tables and their associated odds ratios can be formed for each item. There are potentially 21 of these tables for each item, because one table will be associated with each total number-right score from 0 to 20. In this example, the number-right score is the stratifying variable for the frequency table.

The design of the ECLS-B direct child cognitive assessments administered during the preschool and two kindergarten waves influences the identification of the appropriate stratifying,

or blocking, variable. The cognitive assessment used during the preschool wave included a series of skip/discontinue patterns in the early reading (language and literacy) sections, and second-stage easy or hard mathematics forms were given only to selected children based on performance on a common set of items (i.e., the routing test). The ECLS-B kindergarten assessments also followed an adaptive, two-stage, multiform design. As a result, not all children received the same items or items of the same difficulty, making number-right scores inappropriate for use as stratifying variables. Instead, the IRT ability estimate, theta, was used as the stratifying variable, divided into 41 equally spaced intervals.[44] Accordingly, 41 two-way tables were produced for each item, one for each theta interval. The first dimension of each of the 41 two-way tables is population subgroup (e.g., White children versus Black children), and the other dimension is passing versus failing on a given item. Thus, the question that the M-H procedure addresses is whether members of the reference group (e.g., White children) who have the same total ability estimate as members of the focal group (e.g., Black children) have the same likelihood of passing the item in question. Although the M-H statistic looks at passing rates for two groups while controlling for total score, no assumption need be made about the shape of the total score distribution for either group. In this case, the Chi-square statistic associated with the M-H procedure tests whether the average odds ratio for a test item, aggregated across all 41 score levels, differs from unity (i.e., equal likelihood of passing the item, given the same overall test score).

The M-H procedure has an effect size that is expressed in an odds-ratio metric. Odds-ratios have a minimum value of 0 and a maximum value of positive infinity. Odds-ratios are difficult to interpret because of this range. A more common measure of difficulty is the proportion-correct or p-value. Historically, ETS test developers have worked with a delta metric instead of a p-value to describe item difficulty. To obtain a delta, the proportion correct is converted to a z score via a p-to-z transformation using the inverse of the normal cumulative function, followed by a linear transformation to a metric with a specified mean and standard deviation, such that large values of delta correspond to difficult items, with easy items having small values of delta. Typically, deltas are expressed as integers; p-values are expressed as proportions or percents. ETS has developed a classification scheme for DIF that uses the M-H Delta Difference, or M-H D-DIF, as an effect size for DIF. The M-H D-DIF is an estimate of differences in delta value between a focal group and a reference group. The classification scheme defines a letter code of "A" for negligible DIF, "B" for intermediate DIF, and "C" for large DIF. Items are classified as "A" if either the M-H DIF is not statistically different from zero or if the magnitude is less than one delta unit in absolute value. Items are classified as "C" if M-H DIF both exceeds 1.5 in absolute value and is statistically significantly larger than 1.0 in absolute value. All other items are classified as "B." Items labeled "A" or "B" are considered to have differences that are too small to be important.

[44] The initial estimates of theta in PARSCALE range from -4.0 to +4.0 in intervals of 0.2, resulting in 41 intervals.

The standardized P-DIF procedure is similar in most ways to the M-H method, with the exception that the P-DIF method uses a proportion correct difference metric, while M-H uses a delta difference metric. (The proportion correct metric is defined as the comparison of the proportion correct of the reference and focal groups.) P-DIF has an advantage over M-H for those items in the extremes of the distribution: the P-DIF procedure looks at differences in adjusted proportions of correct item responses, while M-H looks at the log odds ratios. For this reason, the M-H procedure is more susceptible than the P-DIF to a false indication of C-level DIF for items at the extreme values of the difficulty distribution.

The P-DIF uses a weighting function supplied by the focal group to compute the average differences across levels of the matching variable.[45] The focal group supplies specific weighting factors at each score level to weight differences in item performance between the focal and reference groups. In essence, the standardized P-DIF index equals the difference between the observed performance of the focal group (e.g., Black children) on the item and the predicted performance of selected reference group members (e.g., White children) who are matched in ability to those in the focal group. The biggest differences between the M-H D-DIF and the standardized P-DIF estimates are that the standardized P-DIF is easier to understand because its effect size is expressed in a metric that is more intuitive while the M-H D-DIF uses more complex statistics in detecting DIF. The two procedures yield measures that correlate in the high 90s; if discrepancies are observed, the trend is seen for very easy and hard items, items that have little or no impact on the measurement process.

The P-DIF index can range from -1 to +1 (or -100 percent to +100 percent). Positive values indicate that the item favors the focal group, whereas negative values indicate that the item disadvantages the focal group. P-DIF values between -0.05 and +0.05 are considered negligible. Values between -0.10 and -0.05 and between +0.05 and +0.10 are inspected to ensure that no possible effect is overlooked. Items with values outside the -0.10 to +0.10 range are more unusual and are identified as exhibiting DIF with practical significance.

Combining results from both the M-H and P-DIF procedures is advantageous in estimating the existence of statistical DIF. Items with a standardized P-DIF index greater than 10 percent (less than -0.10 or greater than +0.10) *and* with C-level DIF using the M-H method are highly likely to be differentially functioning. Items showing *either* C-level M-H DIF or P-DIF are less likely to be exhibiting statistical DIF but are inspected further. (For example, items in the extremes of the difficulty range may show C-level DIF and not P-DIF. For this particular condition, the item is not considered to be exhibiting differential behavior.)

The fact that an item is identified by DIF procedures does not mean that the item is necessarily unfair to any particular group. DIF procedures are merely statistical screening steps

[45] The reference groups used for the ECLS-B are White, non-Hispanic race/ethnicity, male gender, and for kindergarten enrollment, those enrolled. The former two reference groups adhere to the long-term ETS policy regarding DIF calculations; the latter was added specifically for the ECLS-B.

that indicate that the item is behaving somewhat differently for one or more subgroups. Thus, the formal DIF analysis is the first step in a two-step screening procedure. The second step is a review of the item content for evidence that the item may be measuring some extraneous dimension not consistent with the test framework. Items that exhibit DIF, either in favor of the majority group or against the majority group, are routinely submitted to content analysis by reviewers who were not involved in the development of the test. If the reviewers decide that the item is measuring important content consistent with the test framework and does not contain language or context that would be unfair to a particular group, the item is kept in the test. If the committee finds otherwise, the item is removed from the scoring procedures.

DIF procedures were carried out after each wave of the ECLS-B preschool and kindergarten assessments. The examination of DIF in each subsequent wave included the item data from the prior wave(s). In other words, the kindergarten 2006 wave analysis also included data from preschool wave, and the kindergarten 2007 wave analysis also included data from the kindergarten 2006 and preschool waves.

Items were checked for differential functioning using child's sex, race/ethnicity, and kindergarten enrollment as analysis characteristics. The sex contrast compared males (reference group) with females (focal group). The race/ethnicity contrast groups included White children (reference group) compared with three other racial/ethnic groups of children: Black, Hispanic (any race), and Asian (including Native Hawaiian and Pacific Islander). There were too few American Indian/Alaska Native and multiracial children for DIF statistics to be evaluated separately for these groups, and they are excluded from the DIF analysis altogether. The kindergarten enrollment contrast group was defined as those children "in kindergarten" (reference group) compared with those "not yet in kindergarten" (focal group).[46] (As stated above, all contrasts were controlled on matched ability levels, including the kindergarten enrollment comparison.) Statistics were computed for each item for which the minimum number of required responses, 500 observations for the smaller group, was available. The results of DIF analysis for the kindergarten 2007 wave, which includes data from the kindergarten 2006 and preschool waves, are discussed in detail in chapter 4. It should be mentioned, however, that two early reading items were omitted from construction of the early reading scores due to DIF. No math items were omitted. See section 4.2.4 for more details.

3.4 Development of the Preschool Through Kindergarten Longitudinal Scale

The study of the relationships between children's early childhood experiences through kindergarten entry and their gains in academic skills requires accurate measurements of

[46] Other kindergarten enrollment options were "in first or second grade," "ungraded," or "homeschooled," but included too few children for statistical contrasting. The kindergarten enrollment DIF calculations were performed to determine if the items measured the construct (early reading or mathematics) the same way across the kindergarten and not-yet-in-kindergarten children.

achievement on scales that can be linked across waves. The development of a vertical scale that spans preschool to kindergarten and has optimal measurement properties throughout the achievement range called for multiple test forms that vary in their difficulty (see sections 2.2 and 2.3). The forms are tailored for individuals within a level; however, the overall forms should reflect core curriculum elements for that age or data collection wave. At the same time, there must be overlapping items shared by forms within a wave and across waves. These linking items tie the vertical scale together, both across forms within a wave and across waves. At least 20 percent of the items should overlap across forms and between adjacent waves.

The ECLS-B preschool early reading assessment consisted of a single test form, administered to all children, so item overlap is not relevant for this assessment during preschool. However, the ECLS-B kindergarten early reading assessment followed a multiple-form design, so item overlap can be examined. Table 16 shows the overlap of items across forms within the early reading portion of the ECLS-B kindergarten assessment. The 24 early reading routing items, taken by all children, serve as common items for linking across forms within the kindergarten waves. The second-stage forms consisted of 16, 21, and 27 items, respectively; thus, the 24 routing items alone exceed the 20-percent overlap requirement.

Table 16. Number of early reading items overlapping across ECLS-B kindergarten forms: 2006–07 and 2007–08

Form	Number of unique overlapping items	Percent of unique overlapping items
Levels 0 and 1	5 + 24 routing	52
Levels 1 and 2	7 + 24 routing	48
Levels 0, 1, and 2	1 + 24 routing	29

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections.

Table 17 shows overlap of the ECLS-B kindergarten assessments with forms from other data collections: the ECLS-B preschool national, the ECLS-B kindergarten field test, and the ECLS-K K–1 assessments (the same assessment was used during fall and spring of the kindergarten and first grade year of ECLS-K data collection).

Table 17. Number of ECLS-B kindergarten early reading items overlapping with other assessments: 1998–99, 2005–06, 2006–07, and 2007–08

Assessment	Number of overlapping ECLS-B kindergarten items	Percent of overlapping ECLS-B kindergarten items
ECLS-B preschool national	35	31
ECLS-B kindergarten field test	40	45
ECLS-K K–1 national	50	43
Common to all of the above	14	6

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K), kindergarten data collection, 1998–99, and Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Item overlap can also be examined for the mathematics assessment administered at each wave. As noted in section 2.2, the preschool math assessment included a core set of items,

administered to all children, as well as basal and ceiling secondary forms, administered to children who show low (i.e., basal) or high ability based on the core items. In this design, there was no overlap in item content across the forms, but the core items provided overlap in items across children. The design of the kindergarten mathematics assessment followed a two-stage adaptive design (like the kindergarten reading assessment) using a routing form, administered to all children, with a secondary form administered to children based on their responses to the routing items. Table 18 shows the overlap in items across the kindergarten assessment forms, and table 19 shows the overlap of the ECLS-B kindergarten assessments with forms from other data collections. The 16 routing items, taken by all children, also serve as common items for linking across forms within the kindergarten data collection waves. Similar to the early reading comparison, the second stage forms of the mathematics assessment consisted of 16, 20, and 25 items, respectively; thus, the 16 routing items alone exceed the 20-percent overlap requirement.

Table 18. Number of mathematics items overlapping across ECLS-B kindergarten forms: 2006–07 and 2007–08

Form	Number of overlapping items	Percent of overlapping items
Levels 0 and 1	6 + 16 routing	48
Levels 1 and 2	8 + 16 routing	45
Levels 0, 1, and 2	3 + 16 routing	27

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections.

Table 19. Number of ECLS-B kindergarten mathematics items overlapping with other assessments: 1998–99, 2005–06, 2006–07, and 2007–08

Assessment	Number of overlapping ECLS-B kindergarten items	Percent of overlapping ECLS-B kindergarten items
ECLS-B preschool national	31	42
ECLS-B kindergarten field test	35	54
ECLS-K K–1 national	45	58
Common to all of the above	18	12

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K), kindergarten data collection, 1998–99, and Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Using the general rule listed above, there are ample numbers of items across forms and across assessments to create both horizontal (i.e., within-level and cross-cohort) and vertical (i.e., longitudinal) scales for early reading and mathematics assessments. Although the content and presentation of each of the common items were identical in the versions of the assessments (i.e., preschool, ECLS-K K–1, and kindergarten), it was still possible for the items to function differently. The first step in developing the longitudinal scale was evaluating the functioning of the common items at different time points.

3.4.1 Evaluating Common Items

The longitudinal scales necessary for measuring gain over time were developed by pooling the data from the preschool, kindergarten 2006, and kindergarten 2007 assessments. The

link between the assessment forms used in different waves relied on the presence of common items shared across test forms.

The scale scores for preschool were based on the pool of items used in the test forms administered in that wave. Items were added to the pool when the first kindergarten wave data were collected (kindergarten 2006). Thus, the preschool mathematics scale scores, for example, were estimates based on a pool of 45 items, with the pool expanding to 72 items with the preschool and kindergarten waves combined.[47] When the item pool was expanded, scores were recalibrated for *all three* waves to make longitudinal comparisons possible. The recalibration of the scale score represents the estimated number right on a larger set of items that includes all items in the current wave and all items administered in the previous wave. As a result, the scale score for the *same* child in the *same* wave changes each time a new set of test items is incorporated and the scale on which the score is based is expanded.

Linking score scales across waves required not only overlapping ability distributions (i.e., the high ability levels in the preschool wave overlapped with the low and middle levels of the kindergarten waves), but also overlapping test forms. The longitudinal score scales relied on common items that were present in more than one set of assessment forms. These common items permitted the development of a vertical scale suitable for measuring gains across the final three data collection waves (preschool, kindergarten 2006, and kindergarten 2007). Table 20 shows the number of items in each subject area shared by the assessment forms, as well as the number that appeared in only one set. Within waves, the score scale was developed from items administered to all children within the wave (such as those on the routing forms), as well as smaller numbers of items overlapping two or all three second-stage forms.

Table 20. Counts of common items, unique items, and total items contributing to scale scores for early reading and mathematics: preschool and kindergarten waves

Assessment versions	Early reading	Mathematics
Total item pool	85	71
Common items (total)	18	23
Unique items (total)	67	48
Preschool only	27	18
Kindergarten only	40	30

NOTE: Table includes counts of items included in the scoring for each scale. For each scale, additional items were used in calibration, but excluded from the scoring of the scales because of lack of common-functioning or to better align with the frameworks. Four additional early reading items were used in the calibration of abilities but deleted from the scale scores to bring the content representation into closer alignment with framework specifications. Two additional early reading items were removed for differential item functioning. One additional math item was removed because of an ambiguity in its administration.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

[47] Because the two kindergarten data collection waves used the same instrument, the total item pool was established through the preschool and kindergarten 2006 waves.

The first step in developing the longitudinal scale was evaluating the functioning of the common items at different time points. Although the content and presentation of each of the common items were identical in the two versions of the assessments (preschool and kindergarten), it was still possible for the items to function differently. Of course, it would be expected that performance on the items would improve as children advanced and gained skills, and gains in the probability of a correct answer would be observed. However, the *relative* difficulty of items in the context of the whole assessment should be maintained for the common items used to anchor the scale.

For example, assume an item based on content familiar to children in the preschool and kindergarten waves maintains relative difficulty across waves. This item exhibits the same modeled and observed probability of a correct response at each point on the ability level scale, whether in the preschool or kindergarten waves. In other words, a child at a particular ability level in the preschool wave has the same probability of a correct response to that item as another child in a kindergarten wave with the same ability level. Alternatively, an item X based on content that had not yet been introduced could, in preschool, be the hardest item in the assessment and could be found to be much more difficult than a particular set of computation items Y. By kindergarten, when children could have had practice in the skills tapped by X, it could become much *easier* than the *same* set of Y computations. Such an item, showing a large difference in *relative* difficulty over time, should not be treated as a common item for the purpose of creating a longitudinal scale for estimating gains.

To assess the common *functioning* of the overlapping early reading and mathematics items, preliminary estimates of IRT item and ability parameters were obtained, using all items in the preschool and kindergarten assessment forms. For this purpose, each common item was initially assumed to be common functioning, and then this assumption was tested as follows. Responses for each of the common items were pooled for all waves, and a single set of item parameters was estimated for each. Then the *actual* performance on each of the common items individually in each wave was compared with performance *predicted* by the IRT item and ability parameters to identify discrepancies that would indicate differential functioning for any items.

Comparisons were made between the actual and predicted proportion correct for each of the early reading and mathematics items used in more than one assessment version, based on the children who answered each of the items in each wave of data collection. Note that the comparisons of observed versus predicted proportion correct for each question can be carried out *only for children who answered the question*. Many questions appeared in only one or two second-stage forms within a wave, or after a discontinue point. Thus, some of the items were answered by only a subset of children tested in each wave.

For the majority of the items, the difference between the observed and predicted proportion correct was very small, indicating common functioning of the items across time periods and good fit to the IRT model (see appendix B). The differences in proportion correct for

the reading items were, on average, 0.01, 0.02, and 0.03 for the preschool, kindergarten 2006, and kindergarten 2007 rounds, respectively. The largest differences in proportion correct were observed in the kindergarten 2007 round comparison, for items with only a few hundred observations or for those with low discrimination. The latter was also observed in the preschool and kindergarten 2006 rounds. In math, the average differences were 0.01 for each round, with the largest differences observed in items with low discrimination or low counts. Eight items common to the preschool and kindergarten reading and seven common to the math assessments had sufficiently large discrepancies between observed and predicted percent correct values to warrant separate calibration, based on the fit of the model to the data, and confirmed by the differences in proportion correct, which on average were more than twice as large as the greatest differences for the common-functioning items. These 15 items were deleted from the common item list used for anchoring the scales but retained for each (preschool and kindergarten) assessment form, with separate sets of item parameters. These items were essentially used as separate, distinct items, unlike the common-functioning anchor items, that shared sets of item parameters across waves.

3.4.2 Item Response Theory Evaluation and Scoring

IRT calibration was carried out using the PARSCALE program as described in section 3.2.2. Certain items were deleted from the item pool for calibration and scoring because of DIF for a population subgroup or for administration issues (see sections 4.2.4 and 4.3.4). The estimation of item parameters and child abilities was based on the remaining common and unique items that appeared on all preschool and kindergarten forms of the early reading and mathematics assessments. For each item, the IRT calibration resulted in a set of three item parameters that define a logistic function associated with the item. The height of the function at any point along an ability range corresponds to the estimated probability of a correct answer on the item for a person at that ability level (as shown in figure 5).

Each of the waves of data collection—preschool, kindergarten 2006, and kindergarten 2007—was treated as a separate subpopulation with its own ability distribution for the purpose of IRT calibration. This treatment, which is a feature of PARSCALE and other Bayesian approaches to IRT, provides for an empirically based shrinkage toward subpopulation means for extreme ability estimates, low and high. This shrinkage, which was discussed earlier in section 3.2.2, is particularly important for a longitudinal study, where the focus is on measuring gain and it is important to avoid floor and ceiling effects.

IRT-based scale scores were derived from the IRT item parameters and ability estimates. As described in sections 3.2.1 and 4.1, and illustrated in equation (3.1) in section 3.2.1, the set of three parameters for each item defines a logistic function corresponding to the probability of a correct answer for a test taker with a given ability level. These probabilities are summed over all items to get a scale score representing an estimate of the number of items the child would have answered correctly in each domain.

At each time point (preschool, kindergarten 2006, and kindergarten 2007), the ability estimates were used in combination with the item parameters to generate a probability of a correct response for each item, summed over all items in each domain, for each wave. For example, a child who was tested at all three waves (preschool, kindergarten 2006, and kindergarten 2007) would have three ability estimates and the associated scores for each wave.

The probabilities are summed over all assessment items, excluding those that were not scoreable. Items that were not scoreable include items that exhibited DIF and items that were dropped because of administration problems. Certain items also were not scoreable because they were not common functioning. For the eight early reading and seven mathematics items that were found to not be common functioning across waves of data (as described in section 3.4.1), the fit of the model to the actual distribution of results for each item from each separate wave was examined to determine if the IRT calibration best modeled data from a particular wave. For example, if the IRT model fit the preschool data better than the data in the subsequent kindergarten waves, the item parameters from the preschool version of the item were used in scoring. Conversely, if the IRT model fit the kindergarten data better than the data from the preschool wave, the item parameters from the kindergarten version of the item were used in scoring. Thus, although the item parameters for items like these were calibrated and used in estimating abilities for all waves, the item parameters from *either* the preschool or kindergarten waves were used in scoring. If the fit of the model to the data was not better for any individual wave, the item was retained in the calibration and ability estimation, but dropped from scoring. Four early reading items were calibrated but dropped from scoring. No math items were dropped.

3.5 Evaluating the Preschool–Kindergarten Longitudinal Scale

This section addresses the issue of the validity of the score scales as measures of child achievement and growth between preschool and kindergarten. The validity issue is examined from several perspectives:

Do the tests measure the right content?

Is the difficulty of the tests suitable for children's ability levels?

Do the scores constitute a cohesive scale suitable for longitudinal measurement?

What is the correlation of the same construct across waves (e.g., preschool early reading with kindergarten early reading)? What is the correlation of different constructs within a wave (e.g., preschool early reading with preschool early math)?

3.5.1 Do the Tests Measure the Right Content?

Evidence of the appropriateness of the tests' content can be obtained from two sources: expert judgments and psychometric results. Chapter 2 describes the design of the tests and development of test frameworks. Curriculum experts provided input with respect to cognitive skills that are both typically taught and developmentally important. Test frameworks in each domain were developed accordingly, and test items in each set of assessments were selected to

conform as closely as possible to framework specifications. Field test item pools and proposed final form item selections were reviewed by experts, and the content and presentation of items were modified in response to their recommendations.

A psychometric perspective on the appropriateness of test content included a review of the proportion correct for each item across waves. IRT calibration allows the estimation of performance on each item for *all* waves, even waves in which the item was not used. The match of assessment forms to estimated performance gains (i.e., no significant floor or ceiling effects in any wave in early reading or mathematics) suggests that the content of the tests reflected what children had been learning during the intervening time period.

3.5.2 Is the Difficulty of the Tests Suitable for Children's Ability Levels?

Chapter 2 describes the development of adaptive tests in each subject area for the preschool and kindergarten waves. The adaptive tests were designed to maximize reliability within the available testing time by matching test difficulty to children's ability level while minimizing frustration or boredom that could occur if children received tests that were much too difficult, much too easy, or much too long. Separate assessment packages for preschool and kindergarten focused on items of appropriate difficulty for the waves in which they were administered, while containing enough overlapping items to support the longitudinal scale. Evidence that the tests contained items that were of appropriate difficulty for both the individual children taking them and, in the aggregate, for the waves in which they were administered, can be found in analysis of the test data, specifically in the analysis of floor and ceiling effects. This is particularly important in a longitudinal study, where score scales with floor and ceiling effects can attenuate measurement of gain for the lowest and highest achieving children.

Chapter 4 reviews the operating characteristics of the ECLS-B assessment forms, including the percentages of children with below-chance (to examine presence of a floor effect) and near-perfect (to examine presence of a ceiling effect) scores. No floor or ceiling effects were found on the scale scores for the early reading and mathematics tests in any wave; that is, only a negligible number of children had below-chance or near-perfect scores on the combined routing and second-stage items.[48] A ceiling effect *was* seen in the preschool language measure. Items that were too easy for a significant number of preschool children were excluded from the early reading score measure, thus eliminating the ceiling effect on the measure.

These psychometric results showing no significant floor or ceiling effects in any wave in early reading and mathematics (or effects that were not correctable) indicate that the approach of combining appropriate assessment versions across waves plus adaptive forms within a wave was successful in selecting items of appropriate difficulty for the test takers.

[48] As noted in section 4.1, analysis of items on the preschool early reading items (language and literacy) supported the development of separate scores (one for language and one for literacy). However, when performance on the kindergarten 2006 and kindergarten 2007 early reading items was used to recalibrate the scores, the data supported a single early reading score, combining literacy items and receptive vocabulary. The language items (PreLAS and PPVT items) on their own showed little variability and a ceiling effect, prohibiting development of a standalone language score.

3.5.3 Do the Scores Constitute a Cohesive Scale Suitable for Longitudinal Measurement?

Evidence supports the validity of the score scales for longitudinal measurement in two ways. Examination of IRT "a" parameters (see section 3.2.1) suggests that the item pools within each subject are strongly related to a single underlying factor that is consistent across waves from preschool to kindergarten. The differences in actual and predicted proportion correct demonstrate that the IRT model appropriately represents the test data collected in each wave.

If each test taker had answered *all* of the items in the preschool and kindergarten item pools at *all* waves of data collection, it would be possible to measure the cohesiveness of the scale by observing alpha coefficients and item biserials. Because of time constraints, it would have been neither reasonable nor practical to administer the whole item pool to every child at all three waves. The IRT "a" parameters provide the same type of insight into the cohesiveness of a set of test items. As discussed in section 3.2.1, this parameter represents item discrimination, or the ability of an item to discriminate, or separate, children whose ability level is above or below the calibrated difficulty of the item. In other words, the "a" parameters indicate how strongly each item is related to the underlying construct being measured by the test, with values of 1.0 or above indicating a strong relationship. Values above 1.0 for most of the items in a test constitute evidence that there is a strong underlying factor being measured by the test.

Of the 85 items in the early reading scale, 62 have "a" parameter values greater than 1.0. Those items with "a" parameters near 1.0 are related to understanding conventions of print. Those with the lowest "a" parameters are related to listening comprehension and receptive vocabulary items. The listening comprehension items were challenging for most children, while the receptive vocabulary items were generally easy, characteristics that result in low discrimination for these items. Nearly all of the items tapping early reading skills, from simple letter recognition to decoding, have "a" parameters above 1.0. Results for mathematics were similar, with 50 of 71 items having "a" parameters above 1.0. The remaining items were a mixture of item types, either generally easy or challenging for the sampled children. Although a portion of the items in both the early reading and mathematics assessments had "a" parameters less than 1.0, as discussed above, the extremely small differences between observed and predicted proportion correct for virtually all items at both waves support the idea that the IRT model appropriately represents the test data collected in each wave. In addition, the increase in proportion correct over time, and the fact that increases took place given the content and difficulty of the items, provides further evidence that the IRT results appropriately model achievement growth.

3.5.4 Relationship of the Cognitive Test Scores to Scores in Different Waves and Different Subjects

Table 21 shows correlations of scores for tests in the same subject across waves. Note that, while early reading ability at preschool correlates well with kindergarten reading

achievement (correlation = .65), other experiences between preschool and kindergarten presumably have an important influence as well. Measures of family and school circumstances that relate to child achievement are provided in the ECLS-B database. Exploration of the role these variables play in predicting later achievement is beyond the scope of this report, although analysts would of course want to account for them in their analyses.

Table 21. Correlations of IRT theta score across waves, by subject: Assessment years 2005–06, 2006–07, and 2007–08

Subject	Preschool/ kindergarten 2006	Preschool/ kindergarten 2007	Kindergarten 2006/ kindergarten 2007
Early reading	.65	.58	.69
Mathematics	.72	.64	.77

NOTE: Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. The weight corresponds to the latest wave represented by the correlation. Only those cases with a valid weight are included in the table. IRT = item response theory.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Correlations of scores *across* subjects *within* waves are presented in table 22. The relationship between early reading and mathematics achievement was .76 at preschool, .81 at kindergarten 2006, and .77 at kindergarten 2007. The reading and mathematics theta scores correlate well (approximately .75 or higher) for all waves.

Table 22. Correlations of IRT theta score across subjects, by wave: 2005–06, 2006–07, and 2007–08

Wave	Early reading x mathematics
Preschool	.76
Kindergarten 2006	.81
Kindergarten 2007	.77

NOTE: Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table. IRT = item response theory.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Chapter 4
Psychometric Characteristics of the Direct Cognitive Battery

This chapter documents the results of the direct cognitive assessments in the preschool, kindergarten 2006, and kindergarten 2007 waves of the ECLS-B. The types of scores derived from the assessments are described, along with the psychometric characteristics of each. The emphasis in this chapter is on the psychometric characteristics and scores of the preschool and kindergarten assessments.

Section 4.1 describes the types of scores available on the longitudinal data file for the preschool and kindergarten waves. Section 4.2 discusses the early reading assessments for the preschool and kindergarten waves. Section 4.3 describes the math assessments for the preschool and kindergarten waves. In section 4.4, the administration and scoring of the Let's Tell Stories items are described, along with procedures for assessing the reliability of the coding operations. A description of the color knowledge items, which were fielded only in the preschool wave, is provided in section 4.5.

4.1 The Direct Child Cognitive Assessment Scores

The design and development of the direct child assessments used during the preschool and kindergarten waves of data collection was described in chapter 2. This section documents the types of scores available on the longitudinal data file, including overall scale scores and theta scores.

Overall Scale Scores. For early reading and math, overall scale scores based on the full set of assessment items were calculated using item response theory (IRT) procedures. IRT methods make it possible to calculate scores that can be compared regardless of which assortment of items a child received. The IRT scale scores reported here represent estimates of the number of items children would have answered correctly if they had received all of the scored questions in a given content domain. However, the IRT scale scores are not integers because they consist of probabilities of correct answers summed over all items in the pools. Also, scores for different subject areas are not comparable to each other because they are based on different numbers of questions, as well as content that is not necessarily equivalent in difficulty. That is, it would be incorrect to assume that a child performs better in early reading than in mathematics because his or her IRT scale score is higher for early reading than for mathematics.

Theta Scores and Standard Errors of Measurement. Theta scores (derived from the means of the posterior distributions of theta) estimate ability in a particular domain. Theta scores for early reading and mathematics are provided on the data file for the preschool and kindergarten waves. The IRT-based scale scores described above are derived from these theta

scores. The theta scores are calibrated on the same metric as the IRT item-level difficulty parameters. Therefore, the theta scores may be less intuitively interpretable than the scale scores. However, the theta scores tend to be more normally distributed than scale scores because they are not dependent on the item difficulty parameters of the items within the scale score set. For example, the scale scores are on a metric that translates to the summed number of items correct. To calculate the IRT-based overall scale score, the theta is used to produce a probability for each item that a child would have gotten that item correct. Then, these probabilities are summed into a scale score. However, the probability that a child would have gotten an item correct is dependent on the discrimination, difficulty, and guessing parameters of the item in addition to the ability estimate of the child. Therefore, in an item set such as early reading in the kindergarten waves, where many of the items have high difficulty parameters (resulting in low probabilities correct), the scale score tends to be skewed toward low scores. However, the early reading thetas do not exhibit any skewness. Additionally, because preschool scores are recalibrated and the scale extended upward with the addition of the kindergarten items, the recalibrated preschool scale scores skew increasingly toward the low end of the scale. Again, the associated thetas are not skewed.

Standard errors provide a measure of uncertainty of the theta score estimates. Adding and subtracting twice the standard error from the theta score estimates provides an approximate 95 percent confidence interval where the true theta score most likely occurs.

Differences Between the ECLS-B 9-month–Preschool Restricted-Use Data File and the ECLS-B 9-month–Kindergarten 2007 Restricted-Use Data File. As noted in chapter 3, child cognitive assessment scores are developed through IRT modeling that includes all available data at each wave. The pooling of these data leads to more stable estimates. One consequence of this is that scores developed during an earlier data wave are updated to reflect the recalibration of the scale with subsequent item performance data. As a result, even when the item composition of an IRT scale score remains unchanged, the theta and scale scores for any given child may be expected to change as a result of recalibration with additional data. An additional consequence of recalibration is that score composition can change. This was evident in constructing longitudinal scores for early reading, but not for mathematics. While data from the preschool wave alone supported the development of unique scores for the dimensions of language and literacy (i.e., performance on the language-based items varied uniquely from performance on the literacy-based items), once the preschool data were pooled with the kindergarten 2006 data, it was determined that separate language and literacy scores were no longer appropriate. The longitudinal model for the preschool and kindergarten 2006 and 2007 waves supported a unidimensional early reading domain, which reflects children's performance on certain language-based items (receptive language/PPVT items) and literacy items (e.g., conventions of print, letter recognition, understanding of letter-sound relationships, phonological awareness, sight word recognition, understanding words in the context of simple sentences). As a result, the reestimated preschool IRT thetas and resulting scale scores available in the 9-month–

kindergarten 2007 data file replace the preschool scores in the 9-month–preschool file previously released. That is, there is now a single early reading score for the preschool data and there are no longer separate language and literacy scores. Finally, in consultation with psychometricians, IRT-based subscale scores presented on the preschool data file have been dropped from the data set. Because the overall early reading and mathematics scores are each unidimensional, the inclusion of additional subscores based on the overall IRT thetas was determined to add no additional information beyond what is presented through each scale score and theta estimate. The specific variables released with the preschool data that have been dropped or replaced by recalibrated scores on the 9-month–kindergarten 2007 file are X3RECVOC, X3LITSC, X3LITTS, X3LTR, X3PRINT, X3PHONO, X3LTRK, X3PLSS, X3PLAS, X3LANGTH, X3LITTH, X3MTHSC, X3MTHTS, and X3NMBR.

4.2 Early Reading Assessment

As was discussed in chapters 2 and 3, the early reading assessment was adaptive. That is, not every child received every item in the early reading battery. This section presents information on the operating characteristics of the early reading battery (section 4.2.1) and on the early reading scores that are presented on the data file (section 4.2.2).

4.2.1 Early Reading Battery

4.2.1.1 Operating Characteristics for the Preschool Wave

The preschool early reading battery included both language and literacy assessment items which were administered to children as a single assessment. The language portion of the preschool assessment consisted of 10 items from the PreLAS Simon Says subtest, 10 items from the PreLAS Art Show subtest, and a set of 16 vocabulary items selected from the PPVT-Third Edition (PPVT-III; Dunn and Dunn 1997). During administration, each of the three components was discontinued if a child failed to give *any* correct responses to the first five questions (including two practice items). Discontinue rules were employed within each group of items, so the test could proceed to the next component after a child began having difficulty with the harder items within a specific component. As discussed in chapter 2, in addition to providing scoreable data for the language domain, these items were used to assess whether the child possessed sufficient English skills to understand the basic instructions and premises required to be assessed in English. Only those children who correctly answered at least one nonpractice item were administered the rest of the language, literacy, color knowledge, and mathematics items from the direct child assessment.

Table 23 presents sample counts and operating characteristics of the language portion of the preschool assessment. Counts are shown for children with no item response data for the language assessment, children whose English skills precluded participation in the assessments conducted in English, and children who answered some language items but too few for computation of reliable and valid language scores in the preschool wave. Approximately 250

children did not continue the cognitive assessments in English because they did not demonstrate sufficient English fluency. The "scoreable cases: language assessment" line represents children who met the criteria for English fluency and attempted at least 10 items in all of the language sections combined. The "perfect score" and "chance score"[49] percentages are based on scoreable cases for the language assessment (i.e., the denominator is about 8,450 and not the full sample of about 8,750).[50]

Table 23. Language assessment samples and operating characteristics, ECLS-B preschool data collection: 2005–06

Characteristics	*n*	Percent of scoreable cases
Total sample size	8,750	†
Number of children with insufficient English fluency	200	†
Number of children to be assessed in English	8,550	†
No language items or fewer than 10 attempted	100	†
Scoreable cases: language assessment	8,450	100
Perfect score: PreLAS Simon Says	3,350	40
Perfect score: PreLAS Art Show	4,350	52
Perfect score: vocabulary items	50	1
Perfect score: all language items	50	1
Chance score or below: PreLAS Simon Says	350	4
Chance score or below: PreLAS Art Show	150	2
Chance score or below: vocabulary items	850	10
Chance score or below: all language items	100	1

† Not applicable.

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Details may not sum to the total due to rounding. Children not responding correctly to any of the first five items in each section (Simon Says, Art Show, vocabulary), including two practice items in each section, did not receive the rest of the language assessment, the literacy, color knowledge, or mathematics assessments in English. Estimates are based on the children assessed in English. The number of children shown here may not correspond to the unweighted number of preschool wave language cases included in child assessment score statistics because some child cases are excluded due to parent nonresponse and weighted tables include only cases with valid parent respondent weights. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. The PreLAS items were not multiple choice, so guessing was not an option and zero is treated as the chance score. Percentages are unweighted.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

A significant ceiling effect was evidenced since 40 percent of children tested received perfect scores on the Simon Says subtest, and 52 percent tested received perfect scores on the Art Show subtest in the preschool wave. This result is not unexpected, in light of the sample design of the ECLS-B. The sample is not a representative sample of the population of preschool-aged children, which might include a substantial number of recent immigrants, but represents the population of children born in the United States in 2001. Except for a small number of children who may have lived elsewhere for a period of time and then returned to the United States, these children have lived in the United States all their lives. It is not surprising that even children

[49] "Perfect scores" are correct answers to all items administered, and "chance scores" are at the guessing level or below.

[50] For a case to be scoreable, the assessment had to be conducted in English, and the child had to have attempted at least 10 items in the domain being scored (i.e., language, literacy, or math during the preschool wave).

raised in families whose primary language is not English have, by the preschool year, acquired some English fluency through exposure in day care settings, from watching television, or from other activities.

In terms of a ceiling effect, for 8 of the 10 items in each of the two PreLAS sections, at least 86 percent of the preschool sample answered correctly. Of the four items (two from each PreLAS section) with a lower percentage correct (less than 86 percent), two had potential flaws related to item presentation or differential item functioning (DIF), or both (one was ultimately removed from scoring; see section 4.2.2). No such ceiling effect was found for the vocabulary items, with difficulty statistics throughout the range of 29 percent to 80 percent correct, and less than 1 percent of children answering all 16 items correctly. Unlike the PreLAS items, there was no expectation that the vocabulary items would exhibit a ceiling effect.

Floor effects[51] in the language test were minimal (based on the number of children with chance scores or below), with the exception of the vocabulary items, which showed a small effect. This is most likely a result of the relatively easy language screener used for the assessments. Only about 1 percent of the sample was routed out of the English forms, based on scores on the PreLAS language fluency items. Therefore, children with relatively low English fluency could have passed the language screener and been assessed in English. The vocabulary items on the PPVT are more difficult than those on the language screener; therefore, children with borderline English fluency could have had difficulty with the items and guessed at most of them, resulting in a below-chance score.

The literacy portion of the preschool assessment contained 37 items representing the following content areas related to emergent literacy:

letter recognition, in both receptive and expressive modes (8 items);

letter sounds (6 items);

early reading: recognition of simple words (4 items);

phonological awareness (10 items);

knowledge of print conventions (8 items); and

matching words (1 item).

Several skip rules were employed to preclude administration of items that were much too difficult for children:

After four incorrect answers in the letter recognition section (not necessarily consecutive wrong answers), the remaining letter recognition items, as well as the letter sounds and word reading questions, were skipped and the assessment proceeded to the phonological awareness section.

[51] Floor effects are defined by a pooling of scores at the chance level or below, not as zero scores, since even random guessing would produce correct answers for at least some of the multiple-choice items.

After three incorrect answers to the letter sounds items (including one practice item), the remaining letter sound item(s) and the word reading questions were skipped.

If the child was unable to read the first two simple words, the remaining two were skipped.

In the group of print convention items at the end of the assessment, if all of the first four print conventions items were incorrect, the remaining three items in the literacy assessment were skipped and the child proceeded to the next domain.

Table 24 presents sample counts and operating characteristics for the literacy portion of the preschool assessment. Scoreable cases are those children assessed in English with at least 10 literacy items attempted.

Table 24. Literacy assessment samples and operating characteristics, ECLS-B preschool data collection: 2005–06

Characteristics	*n*	Percent of scoreable cases
Number of children to be assessed in English	8,550	†
No literacy items or fewer than 10 attempted	250	†
Scoreable cases	8,300	100
Perfect score: all literacy items	#	#
Chance score or below: all literacy items	850	10

† Not applicable.

Rounds to zero.

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Details may not sum to the total due to rounding. Percentages are unweighted. Scoreable cases are those children assessed in English with at least 10 literacy items attempted. Estimates are based on the children assessed in English. The number of cases shown here may not correspond to the unweighted number of preschool wave literacy cases included in child assessment score statistics because some child cases are excluded due to parent nonresponse and weighted tables include only cases with valid respondent weights. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

There was no ceiling effect for the literacy-based items. Only about 50 children, less than 1 percent of the sample, gave correct answers to at least 34 of the 37 items, with fewer than 50 children with perfect scores. About 10 percent of children scored at the guessing level (chance) or below, indicating that the test was not measuring accurately for children with the lowest level of emergent literacy skills, resulting in a small floor effect. Similar to what was found on the vocabulary section of the language portion of the preschool assessment, this may be a result of the relatively easy language screener. Children with limited English fluency could have passed the language screener and been assessed in English, had difficulty with the literacy items, and guessed at most of them, resulting in a below-chance score. About one-third to one-half of all children assessed were unable to recognize or name the eight selected letters of the alphabet, some of them were able to answer at least some of the questions in the other emergent literacy areas.

As described above, upon analysis of the kindergarten wave data, the separate language and literacy scores from the preschool wave were replaced with a combined early reading score, a metric including performance on the language *and* literacy items administered in the preschool wave and the early reading items administered in the kindergarten 2006 and 2007 waves. As a result of this change, it was possible that children who had at least the minimum of 10 items per domain in the preschool language and literacy assessment may or may not have had the minimum number of items required to develop an early reading score.[52] Approximately 8,350 of the 8,750 sampled cases in the preschool wave responded to 10 or more of the early reading items, were included in the early reading calibration, and have early reading scores on the data file.

4.2.1.2 Operating Characteristics for the Kindergarten Waves

Unlike the design of the preschool assessment (a single form with discontinue rules), the kindergarten early reading assessment used an adaptive two-stage design composed of four forms: a core, or routing, test of 24 items administered to all children, and three alternative supplementary forms. Which of the three supplementary forms was administered to a particular child depended on the number of correct responses the child gave on the routing form. The core test included a range of easy and hard items sufficient to accurately measure the ability of the majority of the sample. As discussed in chapter 2 (see section 2.2.3), tests in which all items are too hard for some children in the sample and too easy for others can result in imprecise estimates of ability from floor or ceiling effects. The supplementary forms were included as part of the assessment to provide sufficient information on children's abilities at all levels and reduce the likelihood of floor or ceiling effects. For children who performed poorly (i.e., answered fewer than 8 correct) on the initial set of 24 items, a supplemental low-difficulty, or basal, form containing 16 easier items was administered to preclude a floor effect. For those children who scored between 8 and 13 correct on the core form, a 21-item middle-difficulty supplementary form was administered. A supplementary high-difficulty, or ceiling, form with 27 items was used to avoid a ceiling effect for children who gave correct answers to at least 14 items of the core form.

The early reading assessments for the kindergarten waves contained items representing the following content areas:

Basic skills (53 items), which include letter recognition, in both receptive and expressive modes; letter sounds; early reading—recognition of simple words; phonological awareness; knowledge of print conventions; and word matching.

[52] For example, a child with only five valid language responses and five valid literacy responses would have been excluded from the separate language and literacy calibrations and scoring, but would be included in the early reading analysis if those combined 10 items were among those on the early reading scale. Conversely, a child with 10 valid language responses and no literacy responses would have a language score but not a literacy score. If all of those valid items were from the PreLAS Simon Says section, the child would not have an early reading score, since none of the PreLAS Simon Says items contributed to that domain.

Initial understanding (10 items), which requires early readers to provide an initial impression or global understanding of what they have read.

Developing interpretation (2 items), which requires early readers to extend their initial impressions to develop a more complete understanding of what was read.

Demonstrating a critical stance (2 items), which requires early readers to demonstrate an understanding of the story.

Vocabulary (7 items), both receptive and expressive.

As in the preschool assessment, discontinue rules were employed in the early reading assessment to preclude administration of items that were much too difficult for a given child. These rules allowed for children to be skipped out of difficult questions of the same type that they had been unable to answer correctly.

The total number of children who were administered the early reading assessment and its operating characteristics is shown in table 25. There was no evidence of floor or ceiling effects (based on the low number of children with chance or perfect scores, respectively) on the early reading assessment during the kindergarten 2006 or kindergarten 2007 waves.

Table 25. Kindergarten early reading assessment samples, by operating characteristics: 2006–07 and 2007–08

	Kindergarten 2006		Kindergarten 2007	
Characteristics	*n*	Percent of scoreable cases	*n*	Percent of scoreable cases
Number of children to be assessed in English	6,850	†	1,900	†
No early reading items or fewer than 10 attempted	50	†	50	†
Scoreable cases	6,800	100	1,850	100
Received low-difficulty supplement	2,200	33	200	11
Received middle-difficulty supplement	2,450	36	650	35
Received high-difficulty supplement	2,150	32	1,000	55
Routing plus high group with perfect score	#	#	#	#
Routing plus low group with chance score or below	#	#	#	1

† Not applicable.

Rounds to zero.

NOTE Sample sizes (*n*) have been rounded to the nearest 50. Details may not sum to the total due to rounding. Estimates are based on the children assessed in English. During the kindergarten 2006 wave, fewer than 50 children were assessed in Spanish; during the kindergarten 2007 wave, all children assessed were assessed in English. The number of cases shown here may not correspond to the unweighted number of cases included in child assessment score statistics because some child cases are excluded due to parent nonresponse and weighted tables include only cases with valid respondent weights. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted. Percents may not sum to 100 due to rounding.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 data collection, 2006–07, and kindergarten 2007 data collection, 2007–08.

4.2.2 Early Reading Scores

Table 26 presents summary statistics for the early reading scores and thetas, using the 85 scored items[53] administered in the preschool and kindergarten waves.

Table 26. Early reading assessment statistics, by score, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Score	*n*	Mean	*SD*	Low	High
Preschool					
IRT scale score (X3RSCR2)	8,300	25.18	10.07	11.65	80.29
Theta (X3RTHR2)	8,300	−0.49	0.74	−2.47	2.60
Standard error of measurement of theta (X3RSER2)	8,300	0.43	0.18	0.27	1.41
Kindergarten 2006					
IRT scale score (X4RSCR2)	6,750	38.60	14.84	12.39	82.48
Theta (X4RTHR2)	6,750	0.33	0.86	−2.11	3.09
Standard error of measurement of theta (X4RSER2)	6,750	0..34	0.14	0.17	1.24
Kindergarten 2007					
IRT scale score (X5RSCR2)	1,850	48.95	13.23	12.40	82.48
Theta (X5RTHR2)	1,850	0.91	0.70	−2.11	3.09
Standard error of measurement of theta (X5RSER2)	1,850	0.27	0.09	0.17	1.24

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. The sample size used here includes only cases with scores and valid analytic weights, and so may not match tables showing data not requiring weights. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table. IRT = item response theory. *SD* = standard deviation.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

4.2.3 Reliability Statistics for Early Reading Scores

Table 27 presents the reliability statistics for the scores of the early reading assessment for the preschool and kindergarten waves. The most appropriate estimate of the reliability of the early reading assessment as a whole is the reliability of the overall IRT ability estimate, theta. This number is based on the variance of repeated estimates of theta and applies to all scores derived from theta, namely, the IRT scale scores. Error variance was estimated as the within-person variance of repeated estimates of theta, averaged over all data cases. The ratio of this number to the total variance (between-person variance of the posterior mean) is the estimated proportion of total variance that is error variance, and 1 minus the proportion is the estimate of true variance that is reported as the reliability of theta. This reliability index differs from the information function primarily in that it is a single estimate for the whole set of scores, rather than estimates evaluated at each score within the possible range of scores. This index is the most appropriate single estimate of the reliability of the assessment as a whole, because it reflects the

[53] Four items were dropped due to DIF. Two of these (two PreLAS items) were not considered for inclusion in constructing the early reading scores, while the other two were included within the pool of possible items in developing the score. Four additional items were dropped because the items were not common functioning across waves.

internal consistency of performance of all items administered and for the full range of variance found in the whole sample. The reliability of theta applies to all of the IRT-based scores, because these scores are nonlinear transformations of the thetas that do not affect rank orderings. In general, the more items a test has, and the greater the variance in the ability of the test takers, the higher the reliability is likely to be. Reliability is a sample-dependent measure of internal consistency of a test and is related to the size of the test. The reliabilities shown in table 27 are typical for this test size.

Table 27. Early reading assessment reliabilities, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Reliability measure	*n*	Reliability
Preschool IRT-based scores	8,350	.84
Kindergarten 2006 IRT-based scores	6,800	.92
Kindergarten 2007 IRT-based scores	1,850	.93

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Estimates are based on all children with IRT-based scores. The sample size *n* may not correspond to the number of children in the data file who have a valid respondent-level analytic weight because the weight is dependent on the presence of parent interview data, and not all children who were assessed had parents who completed an interview. Additionally, the sample size *n* for preschool is not the same as the *n* presented for either language or literacy during the preschool wave due to the differences in item pools and the required minimum of 10 valid responses. The reliability of the IRT-based scores applies to the theta (ability estimate) and the scale scores. It is based on the ratio of error variance (within-child measurement error) to total variance (across the sample). IRT = item response theory.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005-06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

4.2.4 Differential Item Functioning

Section 3.3 explains the DIF procedures used for identifying test items that perform differentially for population subgroups. The Mantel-Haenszel (M-H) and standardized primary item discrepancy index (P-DIF) results agree for the majority of items, with differences in some of the items with a high percent correct, as expected given the nature of the statistical procedures used.

Table 28 summarizes the results of the M-H DIF and P-DIF analyses of the early reading items for all waves. Both C-level DIF and P-DIF against one or more race/ethnicity focal groups was observed for 12 early reading items (two PreLAS Art Show, six receptive vocabulary, two print convention, one sight word, and one initial understanding item). Two letter recognition items and one sight word item were found to favor the focal group. Of those items favoring the reference group, four were dropped from scoring upon review: two PreLAS, one receptive vocabulary, and one sight word item. Upon review, all items favoring the focal group were retained. An additional DIF computation was done during the kindergarten 2006 wave only using a kindergarten contrast. One letter sound and one print convention item exhibited DIF favoring the reference group in this contrast but these items were retained after review. Items showing evidence of DIF were reviewed by content area experts, and items were retained if these reviewers found no apparent bias in the item to favor one group over the contrasting group(s).

Because the matching criterion used for the DIF analysis was the combined score for all sections, items with a tendency to favor one group are necessarily balanced by items favoring the

opposite group (although not necessarily the same number of items). (See section 3.3 for explanations of the DIF procedures used for identifying test items that perform differentially for population subgroups and the decision process for including or excluding DIF items.)

Table 28. Early reading assessment differential item functioning, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Reference group	In kindergarten/ not yet enrolled in kindergarten[1]	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	1	0	9	8	0
Number of DIF items favoring focal group	0	0	3	0	0

[1] DIF for this contrast computed for kindergarten 2006 wave only. There were not enough cases to run this contrast during the kindergarten 2007 wave.

NOTE: The reference group is listed first in each column (e.g., in kindergarten, male, and White), and the focal group is listed second (e.g., not yet enrolled in kindergarten, female, Black, Hispanic, and Asian). Reference group cells do not sum to the total number of DIF items for that wave because some items showed DIF for more than one group. DIF = differential item functioning.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

4.3 Mathematics Assessment

4.3.1 Assessments by Wave

Similar to the early reading assessment, the mathematics assessment also was adaptive. That is, not every child received every item in the mathematics battery. This section presents information on the operating characteristics of the mathematics battery (section 4.3.1) and on the mathematics scores that are presented on the data file (section 4.3.2).

4.3.1.1 Operating Characteristics for the Preschool Wave

The preschool mathematics assessment consisted of three forms: a core, or routing, test of 28 items administered to all children, and two alternative supplementary forms administered only if needed. The core test included a range of easy and hard items sufficient for accurate measurement for the majority of the sample. Tests in which all items are too hard for some children in the sample and too easy for others can result in imprecise estimates of ability from floor or ceiling effects. For children who performed poorly on the initial set of 28 items (i.e., answered 10 or fewer correctly), a supplemental basal form containing 9 easier items was administered to preclude a floor effect. An eight-item supplementary ceiling form was used to avoid a ceiling effect for children who gave correct answers to most or all of the core form (i.e., correctly answered 21 or more items). Children who answered 11 to 20 items correctly on the core test were not administered any supplementary form.

The preschool mathematics assessment included questions in the following content areas:

number sense (10 items);

geometry (9 items);

counting (14 items);

operations (8 items)[54]; and

patterns (4 items).

The items in the basal supplemental form were related primarily to shapes and counting, and the ceiling supplemental form consisted of simple word and number problems, with and without counters or pictures. Skip rules were employed within the assessment to ensure that children would not be asked to answer questions beyond the level of ability they had already demonstrated. Specifically:

If both practice items in the counting section were incorrect, the rest of the section was skipped.

If the first two number recognition items were incorrect, the remaining two items were skipped.

On the ceiling supplemental form, if the practice item and first operational item in the "number sentence" group were incorrect, the next item (which was also the last item in the mathematics assessment) was not administered.[55]

A minor discrepancy between the core routing specifications and implementation during data collection resulted in a small number of borderline children receiving a basal supplementary form when they should not have received any second-stage test, and a small number at the high end *not* receiving the ceiling supplementary form when they should have, based on their scores on the core test. This occurred because a test item within the core assessment that was to have been scored twice was scored only once. In accumulating the number of correct answers needed to decide whether a basal or ceiling form was to be administered, the item (counting 20 circles on a page) was to have been scored right or wrong according to whether the child successfully counted the 20 circles and then scored *again* as a separate item based on whether the child reached at least 10. However, only the first criterion was scored. The routing rule specified that children who answered 10 or fewer items correctly should receive the basal supplementary form. Children who counted to 20 correctly *and* gave exactly 10 other correct answers in the routing section should not have been routed to the basal supplementary form, but they received it because the additional routing point for correctly counting to 10 was not added to their score. This error had only a slight effect on burden (an additional 9 very easy test items administered to 240 children, taking, on average, less than 2 minutes of extra testing time). It had either a small positive or no effect on accuracy of measurement because additional test items may improve the accuracy of ability estimates, depending on the difficulty levels of the items.

Children who scored 21 or higher on the core assessment were to have been routed to the ceiling supplementary form. A total of about 300 children answered exactly 20 other routing items correctly *and* counted at least to 10, but did not receive the additional point because the

[54] One of the nine original operations items was removed from scoring because of a procedural error in how the item was administered during the direct child assessment. This is described further in section 4.3.1.2.

[55] The basal supplemental form contained a variety of item types, all with similar difficulty levels. Thus, skip rules were not appropriate to skip more difficult items or items of a difficult type.

counting item was not double-scored. The extra point would have resulted in a score of 21, the threshold for administering the ceiling supplement. As a result, these children received the 28 items in the routing test but not the additional 8 items in the ceiling supplementary form.

The routing specifications within the core mathematics assessment were designed conservatively, so a small discrepancy at the borderline was predicted to have a negligible effect on scoring. IRT estimates are derived from the whole pattern of right and wrong answers; to estimate ability with precision, it is important to have a mix of both. If the routing section for the 287 children in question had been scored as intended (i.e., with the counting circles item scored twice), these children would have had 21 questions right, rather than 20 as scored. A score of 21 out of 28 (the total possible correct on the core test) has a sufficient number of wrong answers (7, unless some items were omitted) to provide accurate estimates of ability. The situation would have been much more serious if, for example, children who gave the right answers for all or nearly all of the 28 questions on the routing test did not receive the supplemental form. A cross-tabulation of number of items answered by number correct was examined for evidence of floor or ceiling effects: children scoring below the chance level on the sets of items they received, or children who answered all or nearly all questions correctly. Small numbers of children were found for both conditions, but in neither case was the discrepancy in routing responsible. It can be concluded, then, that this error in routing had a negligible effect on scoring for those approximately 300 children who failed to be routed to the ceiling items.

4.3.1.2 Operating Characteristics for the Kindergarten Waves

The mathematics assessment for both kindergarten waves was administered like the early reading assessment, first with a core, or routing, test of 17 items administered to all children, and then 1 of 3 alternative supplementary forms administered based on the number of correct responses on the routing form. A supplemental basal form was administered to children who performed poorly (i.e., fewer than 6 correct) on the initial set of 17 items to preclude a floor effect. For those children who scored between 6 and 12 correct on the core form, a 20-item middle-difficulty supplementary form was administered. For those children who gave correct answers to at least 13 items of the core form, a supplementary high-difficulty, or ceiling, form with 25 items was administered to avoid a ceiling effect. The items in the basal supplemental form were related primarily to shapes and counting, while the ceiling supplemental form comprised simple word and number problems, and pattern matching items, with and without pictures. The middle-difficulty supplementary form included a combination of these items types, overlapping with the more difficult items from the basal form, and the easier items from the ceiling form. Skip rules were employed within the middle and ceiling supplementary forms to ensure that children would not be asked to answer questions beyond the level of ability they had already demonstrated.

The kindergarten mathematics assessment included questions in the following content areas:

number sense, properties, and operations (41 items);

measurement (3 items);

geometry and spatial sense (4 items);

data analysis, statistics, and probability (3 items); and

patterns, algebra, and functions (7 items).

Actual implementation of the kindergarten 2007 assessment resulted in a minor modification of its initial design. During data collection for the kindergarten 2006 wave, a secondary prompt for a number sense, properties, and operations item related to subtraction was found to be potentially misleading. The prompt, included in the preschool and kindergarten 2006 assessments, but removed from the kindergarten 2007 assessment, was presented in instances when the child did not initially respond to the assessment item. The secondary prompt asked children how many items (in the stimulus prompt) they had in all. This prompt was potentially misleading because it could have been construed by the child as referencing the total number to start with ("in all"). Because the computer-assisted personal interview instrument did not capture whether the child response was made before or after this secondary prompt, the number of children affected by the potentially misleading prompt is not known. Thus, a decision was made to exclude the item from the construction of any child assessment scores (both preschool and kindergarten scores). However, because the item was included on the math assessment routing form in kindergarten 2006, its administration in the field meant that some children may have been misrouted to either a higher or lower second-stage form (based on total number of routing items answered correctly) as a result of their answering this question correctly or incorrectly. In the kindergarten 2006 wave, about 14.5 percent of children were routed to a second-stage form based on this item. About 4.5 percent of children were routed to a more difficult second-stage form based on a correct response to this item than they would have been if the item had not been counted as correct. About 10 percent of children were routed to a lower second-stage form based on an incorrect response to this item than they would have been if the item had not been counted as incorrect.

The implications of the removal of this item from scoring, but not routing, were examined closely using several approaches. First, the overlap in ability distributions/item difficulties across forms was confirmed to ensure accurate measurement of children at the cut points (no floor or ceiling effects).[56] Second, the standard deviation of the theta estimates for those children at the cut points was comparable to the average standard deviation of the estimate

[56] The examination of the ability and difficulty distributions at the cut points showed adequate overlap such that children routed to a form other than the one to which they should have been would not have a negative impact on their scores. For example, if the three second-stage forms did not have overlapping difficulty/ability distributions, the measurement of a child's ability at the cut points could have been impacted if the child was misrouted.

for the rest of the sample. Finally, removal of the item from the routing form, and its effect on the measurement of ability levels for the children who were possibly misrouted and those who were routed properly is not statistically significant based on the estimate of the standard error of measurement on the form.[57] Scores are provided for these children, and the cases potentially affected by this error are flagged by the variable C4MA_F1 in the supplementary errata dataset included in appendix E on the longitudinal ECLS-B electronic codebook (ECB) data DVD. The item was removed from the kindergarten 2007 administration.

Tables 29 and 30 display the total number of children administered the mathematics assessment and its operating characteristics for the preschool wave and for the kindergarten 2006 and kindergarten 2007 waves, respectively.

No significant ceiling or floor effects (i.e., low numbers of children with perfect or chance scores) were observed in the preschool mathematics test. Of the highest ability children (i.e., those routed to the ceiling supplementary form), fewer than 50 gave correct answers to 34 or more of the 36 items they received. This represents about one quarter of 1 percent of the tested sample. About 1 percent of the children taking the mathematics test scored at the chance level or below, with fewer than four items correct. There was no evidence of floor or ceiling effects in the kindergarten 2006 or kindergarten 2007 waves.

[57] This standard error of measurement (SEM) check was done on the routing test items using number-right scores and not the IRT calibration. Because the children were routed based on their number-right routing score, the impact in this metric was checked. The SEM was estimated as the sqrt(1-A)*SD, where A=alpha reliability of the routing test and SD=standard deviation of the routing test score. The SEM was calculated for the routing test with and without item MAT045. In both cases, the SEM was approximately 1.5, or equivalent to that of many items. Thus, dropping this single item was within the bounds of the error of measurement and does not introduce any measurable error to the scores.

Table 29. Mathematics assessment samples and operating characteristics, ECLS-B preschool data collection: 2005–06

Characteristics	Number	Percent of scoreable cases
Number of children to be assessed in English	8,550	†
No mathematics items or fewer than 10 attempted	250	†
Scoreable cases	8,300[1]	100
Received first-stage (routing) form only	4,750	57
Received low-difficulty supplement	2,700	32
Received high-difficulty supplement	900	11
Routing plus high group with perfect score	#	#
Routing plus low group with chance score or below	100	1

† Not applicable.

Rounds to zero.

[1] A few children answered enough first-stage items to receive a score, but the test was discontinued and no supplemental form was administered.

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Details may not sum to the total due to rounding. Percentages are unweighted. Estimates are based on the children assessed in English. The number of children shown here may not correspond to the unweighted number of cases included in child assessment score statistics because some child cases are excluded due to parent nonresponse and weighted tables include only cases with valid respondent weights. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Table 30. Kindergarten 2006 and kindergarten 2007 mathematics assessment samples, by operating characteristics: 2006–07 and 2007–08

	Kindergarten 2006		Kindergarten 2007	
Characteristics	*n*	Percent of scoreable cases	*n*	Percent of scoreable cases
Number of children to be assessed in English	6,850	†	1,900	†
No mathematics items or fewer than 10 attempted	50	†	50	†
Scoreable cases	6,850	100	1,850	100
Received low-difficulty supplement	1,350	19	150	8
Received middle-difficulty supplement	4,150	61	1,050	55
Received high-difficulty supplement	1,350	20	700	36
Routing plus high group with perfect score	#	#	#	#
Routing plus low group with chance score or below	#	#	#	#

† Not applicable.

Rounds to zero.

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Details may not sum to the total due to rounding. Percentages are unweighted. Estimates are based on the children assessed in English. The number of cases shown here may not correspond to the unweighted number of cases included in child assessment score statistics because some child cases are excluded due to parent nonresponse and weighted tables include only cases with valid respondent weights. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections.

4.3.2 Mathematics Scores

Table 31 presents summary statistics for the mathematics scale scores and thetas, using the 71 scored items administered in the preschool and kindergarten waves.

Table 31. Mathematics assessment statistics, by score, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Score	*n*	Mean	*SD*	Low	High
Preschool					
IRT scale score (X3MSCR2)	8,250	29.31	9.56	9.83	65.74
Theta (X3MTHR2)	8,250	−0.47	0.78	−2.84	2.38
Standard error of measurement of theta (X3MSER2)	8,250	0.40	0..04	0.33	1.10
Kindergarten 2006					
IRT scale score (X4MSCR2)	6,750	40.40	10.56	11.06	69.69
Theta (X4MTHR2)	6,750	0.38	0.80	−2.42	3.12
Standard error of measurement of theta (X4MSER2)	6,750	0.34	0.06	0.26	0.98
Kindergarten 2007					
IRT scale score (X5MSCR2)	1,850	47.72	9.52	10.85	69.69
Theta (X5MTHR2)	1,850	0.92	0.71	−2.48	3.12
Standard error of measurement of theta (X5MSER2)	1,850	0.32	0.04	0.26	0.82

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. The sample size used here includes only cases with scores and valid analytic weights, and so may not match tables showing data not requiring weights. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table. IRT = item response theory. *SD* = standard deviation.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

4.3.3 Reliability Statistics for Mathematics Scores

Table 32 presents reliability statistics for the scores of the preschool, kindergarten 2006, and kindergarten 2007 mathematics assessments (calculated similarly to the reading reliability statistics). In general, the more items a test has, and the greater the variance in the ability of the test takers, the higher the reliability is likely to be. Reliability is a sample-dependent measure of internal consistency of a test and is related to the size of the test. The reliabilities shown in table 32 are typical for this test size.

Table 32. Mathematics assessment reliabilities, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Reliability measure	*n*	Reliability
Preschool IRT-based scores	8,300	.89
Kindergarten 2006 IRT-based scores	6,850	.92
Kindergarten 2007 IRT-based scores	1,850	.92

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Estimates are based on all children with IRT-based scores. The sample size *n* may not correspond to the number of children in the data file who have a valid respondent-level analytic weight because the weight is dependent on the presence of parent interview data and not all children who were assessed had parents who completed an interview. The reliability of the IRT-based scores applies to the theta (ability estimate) and the scale scores. It is based on the ratio of error variance (within-child measurement error) to total variance (across the sample). The *n* presented in this table is based on the full number of scoreable cases. IRT = item response theory.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

4.3.4 Differential Item Functioning

Table 33 summarizes the results of the DIF analysis of the preschool, kindergarten 2006, and kindergarten 2007 waves combined. Both C-level and P-DIF against one race/ethnicity focal group was observed for only one mathematics item in the geometry section. One counting item exhibited DIF favoring the focal group. In the kindergarten 2006 wave, the DIF contrast for kindergarten enrollment did not result in any items exhibiting DIF.

The items exhibiting DIF against the focal groups are balanced by other C-level DIF and P-DIF items favoring one or more of the focal groups. Upon review of the items, none was removed from scoring since none was determined to exhibit any bias. (See section 3.3 for explanations of the DIF procedures used to identify test items that perform differentially for population subgroups and the decision process for including or excluding DIF items.)

Table 33. Mathematics assessment differential item functioning, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005-06, 2006–07, and 2007–08

Reference group	In kindergarten/ not yet enrolled in kindergarten[1]	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	0	0	1	0
Number of DIF items favoring focal group	0	0	1	0	0

[1] DIF for this contrast computed for kindergarten 2006 wave only. There were not enough cases to run this contrast during the kindergarten 2007 wave.

NOTE: The reference group is listed first in each column (e.g., in kindergarten, male, and White), and the focal group is listed second (e.g., not yet enrolled in kindergarten, female, Black, Hispanic, and Asian). Reference group cells do not sum to the total number of DIF items for that wave because some items showed DIF for more than one group. DIF = differential item functioning.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

4.4 Let's Tell Stories

The ECLS-B preschool and kindergarten direct child cognitive assessments also included two of the three items from the Let's Tell Stories subscale of the PreLAS (Duncan and De Avila

1998).[58] This section describes the administration and scoring procedures for these items (section 4.4.1) and the scores derived from this portion of the assessment (section 4.4.2). The hiring and training of coders and the maintenance of coding reliability are described in chapter 8.

4.4.1 Administration and Scoring

Children were administered two story items from the PreLAS Let's Tell Stories subtest as part of the direct child cognitive assessment: Rainstorm and Butterfly in the preschool wave and Shoemaker and Butterfly in both kindergarten waves. For each of these stories, the field interviewer pointed to a series of pictures while telling the child a scripted story. After each story was completed, the child was asked to retell the story using the pictures as a prompt, if needed. Children's responses were audiotaped and simultaneously handwritten in the Child Assessment Booklet by the field interviewer to assist later scoring by coders. Upon completion of the child assessments, the audiotape recording from the PreLAS subset and the Child Assessment Booklet were sent to RTI via overnight express mail for receipt and processing. The audiotapes and transcripts in the Child Assessment Booklet were then scored in a secured coding room in a centralized location by child assessment coders specifically hired for that purpose. The transcripts were completed, modified, or corrected as necessary by the coding staff. A standardized scoring sheet was used to document problems encountered in the coders' review of the tapes (e.g., interviewer protocol problems, home environment problems, audiotape quality) and to record scores for the two stories. Examples of problems documented included incorrect labeling of the audiotape, receipt of damaged or incomplete audiotapes, or the field interviewers (FIs) not providing an explanation when a child did not respond to a specific item. Because coding was ongoing during data collection, feedback based on this information was forwarded to specific FIs on an as-needed basis so that they could improve their administration of the items in the field. Periodic data quality e-mails and newsletters also were sent to the entire field staff to reinforce the proper administration of all the items centrally coded (PreLAS and fine motor items). For example, interviewers were reminded about placing the recorder near the child, limiting background noise, and other procedures. Chapter 8 provides a description of coder training and reliability.

4.4.2 Let's Tell Stories Scores

As discussed in chapter 8, coders were trained to use the holistic scoring system provided in the PreLAS scoring manual. Using this system, scores ranged from 0 to 5, and a score of -9 was used to indicate that a given response was not codeable and, consequently, is missing in the data file (and not included in statistical calculations). As shown below, these scores describe the degree to which the child was able to construct a grammatically correct, coherent story:

[58] During each wave, two of the three story items were included. For the kindergarten waves, one of the stories used during preschool was replaced with the story not used during the preschool wave.

0 = No response (includes "I don't know"), or no response in English.
1 = Short, isolated phrases; at least one word in English.
2 = Disconnected thoughts, at least one sentence, many grammatical errors.
3 = Recognizable story line, limited detail, grammatical errors.
4 = A recognizable version of a story in coherent, fluent sentences.
5 = Articulate, detailed sentences, vivid vocabulary, and complex constructions.
-9 = Not ascertained or not codeable.

Scores for each story, as well as the average score across both stories, are included in the data file. Average scores are only reported for cases with valid scores on both stories scored during each wave (if one or both cases do not have valid scores, the average is set to missing [-9]). Table 34 presents means and standard deviations for each of the Let's Tell Stories scores during the preschool, kindergarten 2006, and kindergarten 2007 data collections.

It is important to note that the two stories administered during the kindergarten waves were not the same two stories administered during the preschool wave of data collection. Analysts will have to determine whether item-level scores or the mathematical average are most appropriate when looking at change in scores over time.

Table 34. Descriptive statistics for Let's Tell Stories items, by variable, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

		Preschool			Kindergarten 2006			Kindergarten 2007		
Variable name	Variable label	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
C3PLTSR	Lets Tell Stories—Rainstorm	8,250	2.26	1.09	†	†	†	†	†	†
C*PLTSS	Lets Tell Stories—Shoemaker	†	†	†	6,700	3.30	0.96	1,850	3.47	0.80
C*PLTSB	Lets Tell Stories—Butterfly	8,200	2.53	1.13	6,700	3.42	0.92	1,850	3.60	0.72
X*EXPLNG	Average Let's Tell Stories score	8,200	2.40	1.02	6,650	3.37	0.87	1,850	3.54	0.69

† Not applicable.

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Where variable names and labels are provided, the data wave is not specified. For wave-specific variable names, replace the asterisk (*) with the number 3 for the preschool wave, 4 for the kindergarten 2006 wave, and 5 for the kindergarten 2007 wave. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

4.5 Color Knowledge

During the preschool wave only, a color knowledge test was administered that asked children to name the colors of the 5 teddy bears (out of 10 pictured) indicated by the assessor, with each correct answer receiving 2 points. For all colors the child could not initially name, the assessor asked, "Can you find the [blue] bear?" Children received 1 point per correct answer in

this receptive mode. The color knowledge test was administered along with the preschool mathematics assessment, and a separate score for it is provided on the data file. The color knowledge number-right score is reported only for children who answered items related to at least three of the five colors presented in the set. Table 35 presents the alpha reliability statistic, which indicates how well a set of items measures a single, unidimensional latent construct, in this case, color knowledge. table 36 presents the mean, standard deviation, and possible range of values for the color knowledge score of the preschool assessment.

Table 35. Color knowledge test reliabilities, ECLS-B preschool data collection: 2005–06

Reliability measure	*n*	Number of items	Reliability
Color knowledge test	8,400	10	.82

NOTE: Sample sizes (*n*) have been rounded to the nearest 50.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Table 36. Color knowledge test score statistics, ECLS-B preschool data collection: 2005–06

Score	*n*	Mean	*SD*	Low	High
Color knowledge (X3COLOR)	8,350	8.69	2.34	0	10

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data. The *n* presented in this table is based on the full number of scoreable cases with a valid value for W3R0. *SD* = standard deviation.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Chapter 5
Physical Measures and Fine and Gross Motor Assessments

In addition to direct measures of children's cognitive functioning, the ECLS-B includes direct assessments of children's physical growth and motor abilities. Physical development and motor skill development have been measured during each wave of the ECLS-B. With respect to physical measures, height, weight, and middle upper arm circumference (MUAC) were obtained during the preschool, kindergarten 2006, and kindergarten 2007 waves for all study children, and head circumference was measured for children born at very low birth weight (1,500 grams or less), consistent with the 9-month and 2-year collections. The measurements for child height and weight were used to calculate the child's body mass index (BMI) at each of these data waves.

With respect to motor abilities, during the 9-month and 2-year waves of data collection, the motor skills subtest of the Bayley Short Form-Research Edition (BSF-R) was used to assess motor skills. By the preschool data wave, this assessment was no longer age appropriate for children participating in the ECLS-B and, therefore, was replaced with a set of fine and gross motor items from the Early Screening Inventory–Revised (ESI-R; Meisels et al 1997),[59] the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), the Bruininks-Oseretsky Test of Motor Proficiency (Bruininks 1978), and the Movement Assessment Battery for Children (Henderson and Sugden 1992). This new set of items comprises the ECLS-B psychomotor battery for the preschool and kindergarten waves.

This chapter begins with a description of the physical measurements collected during the preschool, kindergarten 2006, and kindergarten 2007 waves of the ECLS-B, how they were obtained in the field, basic descriptive statistics at each wave, and a discussion of methods used to enhance data quality. It also describes the development of the ECLS-B psychomotor battery administered during the preschool and kindergarten waves, as well as the specific measures of children's fine and gross motor skills used. A discussion of the coding scheme applied centrally at RTI for some of the fine motor items, and the approaches taken to ensure reliability of the coded items in the preschool and kindergarten waves, is provided in chapter 8.

5.1 Physical Measures

The physical measurement battery used for the preschool and kindergarten waves of the ECLS-B is generally consistent with data collected during the 9-month and 2-year waves, although as described below, some modifications in the methods used to obtain measurements and enhance their accuracy were implemented across waves. Section 5.1.1 describes the item

[59] The ESI-R consists of two separate versions, the Early Screening Inventory–Preschool (ESI-P), for children aged 3 to 4½ years and the Early Screening Inventory–Kindergarten (ESI-K), for children aged 4½ to 6 years. Items were taken from both versions of the ESI-R for the purposes of the ECLS-B.

development process and general procedures for obtaining the measurements, while sections 5.1.2 through 5.1.6 describe the physical measurements in detail. Section 5.1.7 presents summary statistics on the physical measures obtained during the preschool and kindergarten waves, and a discussion of the reliability of these measures at each wave is provided in section 5.1.8.

5.1.1 Item Development and General Data Collection Methods

Physical measurements collected during the preschool and kindergarten waves also were taken during the 9-month and 2-year waves. (See the ECLS-B 9-Month–2-Year Data File User's Manual [Nord et al. 2006] for details on the measurement process in these earlier waves.) The ECLS-B collected information on height, weight, MUAC, and head circumference to support researchers' interest in general health and physical development. Data on height and weight supported calculation of the children's BMI. The procedures for obtaining these measurements were adapted from the protocol for the National Health and Nutrition Examination Survey (NHANES) and were used in previous data collection waves (see http://www.cdc.gov/nchs/data/nhanes/meccomp.pdf for an overview of this protocol).

All physical measurements were recorded in the Child Assessment Booklet. During the preschool wave, field interviewers (FIs) collected each measurement two times, and recorded these two values in the Child Assessment Booklet. Before leaving the child's home, the FI also entered these values into the computer-assisted personal interviewing (CAPI) instrument. For the kindergarten waves, this procedure was modified so that each measurement was taken three times and recorded in the Child Assessment Booklet. The interviewer then entered the two closest values for each physical characteristic assessed[60] into the CAPI instrument.

Steps were taken in the field to ensure the validity of the physical measurements by minimizing the likelihood of errors in both measurement and data entry. This was done in two ways: with hard range checks and with soft range checks. To prevent interviewers from recording errors, the CAPI instrument questioned values that were outside an expected range. This expected range was based on growth curves available from the National Center for Health Statistics (see http://www.cdc.gov/growthcharts/). The minimum and maximum possible expected scores for the physical measurements were programmed in the instrument to set absolute limits for the values that FIs could enter for each assessment (called "hard range checks"). Because the FI could not enter a value outside of the hard range,[61] any observed values outside of the hard check range were entered as the minimum or maximum value allowed.[62] In

[60] When assessing MUAC, the interviewer took a series of measurements and consequently entered the two sets of measurements corresponding to the two closest MUAC measurements.

[61] During training, FIs were instructed that they had to enter data within the programmed hard range; if they recorded a value outside of that range in the booklet, they were instructed to enter the minimum or maximum value, as appropriate, instead of a value outside of the range.

[62] The CAPI instrument programming was modified during the kindergarten 2007 wave to allow FIs to enter a numeric value for "not administered" (either 96 or 99) which would be accepted even though it was outside of the hard range check. These values were then identified as not administered in the data set. Prior to kindergarten 2007, FIs used a function key on the laptop to record when an item was not administered.

addition, narrower ranges were programmed into the CAPI instrument; these ranges reflected the expected values for the majority of the study children (i.e., the measurements of an average 4- or 5-year-old). Thus, when entering the data from the Child Assessment Booklet into the CAPI instrument, FIs were prompted to confirm measurements if they fell outside of this average expected range (called "soft range checks"), but did not violate hard range checks. This prompt acted as a check against values that approached the allowable minimum or maximum values but were still expected to be less common, according to growth curves available from the National Center for Health Statistics. This prompt allowed interviewers to check their entries and, if necessary, re-do the measurements before leaving the child's home. Hard range checks were employed during the preschool, kindergarten 2006, and kindergarten 2007 waves, with soft range checks also used during the kindergarten waves of data collection.

The Child CAPI instrument also prompted the FI to confirm a pair of measurements when they differed by more than a specified amount. The CAPI instrument allowed only two values to be entered, so when the difference between the two values was greater than a predetermined allowable amount, the FI was prompted to check the data he or she had entered and correct any errors in the measurement values, or bypass the edit check, thereby confirming the entered values.[63] Table 37 shows the values for each measurement's hard and soft range checks and the allowable difference between the two measures. For some items, the ranges were expanded across data waves to allow for an expected amount of growth from one year to the next. For all physical measurements, each of the two values entered into CAPI, as well as their mathematical average, are included on the data file.

Table 37. Hard and soft range check values and allowable differences for physical measurements, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

	Preschool		Kindergarten 2006		Kindergarten 2007		All waves
Measurement	Hard range	Soft range	Hard range	Soft range	Hard range	Soft range	Allowable difference
Respondent's weight (kg)	35–137	†	35–137	50–100	35–137	50–100	3 kg
Child's weight (kg)	10–50	†	10–65	15–25	10–65	15–30	3 kg
Child's height (cm)	85–120	†	85–150	100–115	85–160	100–122	5 cm
Child's middle upper arm circumference (cm)	11–40	†	11–50	11–45	11–50	11–45	3 cm
Child's head circumference (cm)	42–60	†	42–72	47–60	42–72	47–60	4 cm

† Not applicable. There were no soft checks programmed into the CAPI instrument used during the preschool wave.

NOTE: The data collection protocol called for measuring the parent respondent's weight. These data are retained on the data file only for those respondents who were identified as the child's biological mother in the base year.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005-06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

[63] There were three instances of paired scores differing by the expected value, all were for respondent weight.

5.1.2 Respondent Weight

The first physical measurement taken was the weight of the parent respondent. Respondents stepped onto a SECA® scale, which displayed weight in kilograms.[64] Respondents were asked to remove their shoes and, if appropriate, jackets or heavy outerwear before stepping onto the scale for measurement. Three independent measures were taken and recorded in the Child Assessment Booklet. The FI then entered the two closest measurements into the child CAPI instrument. In most of these cases (96.0 percent during preschool, 94.8 percent in kindergarten 2006, and 92.9 percent in kindergarten 2007), the weight data collected were from the child's biological or nonbiological mother/female guardian. If the parent respondent was the child's biological mother and the 9-month parent respondent, these two weight measurements were averaged to calculate the biological mother's average weight (X*BIOMWT).[65] FIs indicated whether the respondent on this item was pregnant or if some other adult was weighed. As described in section 5.1.1, the CAPI instrument provided a hard range for these values, resulting in some pooling of data at the maximum value (137 kg) during each wave: 0.68 percent (unweighted) of cases during preschool, 0.19 percent (unweighted) of cases during kindergarten 2006, and 0.12 percent (unweighted) of cases during kindergarten 2007.

5.1.3 Child Weight

To obtain the child's weight, the FI instructed the child to stand unassisted on the SECA® scale, as demonstrated by the parent respondent. Children were asked to remove shoes and (if appropriate) jackets and heavy outerwear. Weight was displayed in kilograms. As with the parent's weight, multiple measurements were taken at each wave, with the child stepping off of the scale to allow it to reset to 0.0 kg between measurements. For each wave, the two measurements recorded in the CAPI instrument and their mathematical average are provided on the data file. If only one measurement was obtained, it was also saved as the average. There was less than 0.02 percent of cases where one of the two child weight measures was at the minimum value (10 kg) and no cases at the maximum value (50 kg) during the preschool wave. There were no cases recorded with values at either the minimum or maximum value during the kindergarten 2006 or kindergarten 2007 waves.

5.1.4 Child Height

Consistent with the 2-year data collection,[66] child height was obtained using a SECA® portable stadiometer[67] in the preschool and kindergarten waves. The child stood erect at the base

[64] Weights were obtained using the SECA® Model 840 scale. See https://www.seca-online.com/ for technical specifications.
[65] Where variable names are provided, the data wave is not specified. Analysts looking for wave-specific variable names should replace the asterisk (*) with the number 3 for the preschool wave, with the number 4 for the kindergarten 2006 wave, and with the number 5 for the kindergarten 2007 wave. This variable naming convention is used throughout the user's manual in instances where variable names differ only by the wave indicator.
[66] Note that during the 9-month wave, data were collected on child length. Starting with the 2-year data wave, child height was collected instead.
[67] Height was obtained using a SECA model 214 portable stadiometer (see https://www.seca-online.com/ for technical specifications).

of the stadiometer, with his or her head in the correct position—head upright, facing away from the stadiometer. Then, a crown piece was lowered, and the child's height was measured in centimeters. As described in section 5.1.1, multiple measurements were taken at each wave, with the child stepping off the stadiometer between measurements. The two measurements entered into CAPI and their mathematical average are provided on the data file for each wave. If only one measurement was obtained, it was also saved as the average.

At each wave of data collection, the hard range provided by the CAPI instrument resulted in some pooling of data at the minimum and maximum values. During the preschool wave, the height measurement was at the minimum value (85 cm) for 0.12 percent (unweighted) of cases and at the maximum value (120 cm) for 0.24 percent (unweighted) of cases. During the kindergarten 2006 wave, a small number (0.03 percent, unweighted) of completed cases were at the minimum value (85 cm), and 0.06 percent (unweighted) of completed cases were at the maximum value (150 cm). During the kindergarten 2007 wave, there were no cases at the minimum, and 0.11 percent (unweighted) were at the maximum value (160 cm).

As was done during the 2-year wave, each child's average height and weight were used to calculate the child's BMI, based on the formula found on the Centers for Disease Control and Prevention website (http://www.cdc.gov/bmi):

$$\text{BMI} = (\text{weight (kg)} / [\text{height (cm)}]^2) \times 1000 \qquad (5.1)$$

5.1.5 Child Middle Upper Arm Circumference

The child's MUAC is an indicator of nutritional status and general physical growth. Procedures for measuring the MUAC were adopted from the protocols used in the NHANES. To obtain the MUAC, the FI measured the length of the child's upper arm and found the midpoint.[68] Next, the FI looped a measuring tape around the child's upper arm and tightened it at the midpoint. A measurement in centimeters was then taken. As described in section 5.1.1, each set of measurements was taken up to three times, with the FI repeating all steps in the process for each measurement.[69] The upper arm length, midpoint, and MUAC for the two measurements entered into the CAPI instrument (the two measurements that were most alike were entered), as well as the mathematical average of the two MUAC values, are included on the data file. If only one MUAC measurement was obtained, it was also saved as the average.

At each wave of data collection, the hard range provided by the CAPI instrument resulted in some pooling of data at the minimum value (11 cm) during each wave: 0.49 percent (unweighted) of cases during preschool, 0.98 percent (unweighted) of cases during kindergarten

[68] The protocol specifies that these measures be taken while the child is standing; however, for children who could not stand or were not able to stand still while the measurements were taken, FIs were instructed to take these measurements while the child was seated on the parent's (or other adult's) lap.

[69] The protocol for obtaining the physical measurements during the preschool wave instructed interviewers to obtain each measurement twice. To improve accuracy, the protocol was modified for the two kindergarten waves to instruct the interviewers to obtain each measurement three times, with the two closest in value being entered into the CAPI instrument.

2006, and 0.37 percent (unweighted) of cases during kindergarten 2007. No cases during any wave were at the maximum value.

5.1.6 Child Head Circumference

Head circumference has been used as a proxy indicator of brain growth and development. It is particularly important for children born at risk. Head circumference was obtained from only those children who were born with very low birth weight (less than 1,500 grams). Head circumference was obtained following procedures used during prior waves of the ECLS-B, which were adapted from NHANES. (See http://www.cdc.gov/nchs/data/nhanes/meccomp.pdf for more details on these methods.) To obtain head circumference, the FI asked the parent respondent to remove any braids or hair ornaments from the child's hair. Then, with the child sitting in the mother's lap, the FI looped the retractable tape measure around the child's head, just above the brow and around the largest diameter in back. As described in section 5.1.1, multiple, independent measurements were taken, and two measurements were entered into the Child CAPI Instrument before the FI left the child's home. These two values and their mathematical average are on the data file. If only one measurement was obtained, it was also saved as the average.

At each wave of data collection, the hard range provided by the CAPI resulted in some pooling of data at the minimum value (42 cm). During the preschool wave, 0.24 percent (unweighted) of cases were at the minimum value. During the kindergarten 2006 wave, less than 0.29 percent (unweighted) of completed cases were at the minimum value. During the kindergarten 2007 wave, 0.41 percent (unweighted) of cases were at the minimum value. No cases had values at the maximum of the range during any wave.

5.1.7 Physical Measurement Scores

As described in sections 5.1.2 through 5.1.6, the mathematical average of the two values entered into the CAPI instrument was computed to provide a single estimate for each physical measurement. If only a single measurement was obtained, it was also saved as the average. The average values (i.e., composite scores), as well as each of the measurements recorded in CAPI, are included on the data file. The child's BMI was calculated from the average value for that child's height and weight (X*CHHGHT and X*CHWGHT).[70] The weighted means, standard deviations, and value ranges for these composite variables are shown in table 38.

5.1.8 Reliability of Physical Measurements

Because each of the physical measurements was taken multiple times, with two measurements entered into CAPI, it is possible to obtain an estimate of measurement reliability

[70] For wave-specific variable names, replace the asterisk (*) with the number 3 for the preschool wave, the number 4 for the kindergarten 2006 wave, and the number 5 for the kindergarten 2007 wave.

by examining the correlation between measurements.[71] The composite variable names, labels, and correlations between the two measurements are shown in table 39.

[71] Note that during the preschool wave each measurement was taken only twice; during the two kindergarten waves the measurements were taken three times, with the two closest measures entered into CAPI.

Table 38. Summary statistics for physical measurements by variable, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

		Preschool				Kindergarten 2006				Kindergarten 2007			
Variable name	Variable label	*n*	Mean	*SD*	Range of values	*n*	Mean	*SD*	Range of values	*n*	Mean	*SD*	Range of values
X*BIOMWT	X* biological mother's weight (kg)	7,550	74.96	18.73	37.3–137.0	5,800	75.30	18.35	36.9–137.0	1,500	77.23	18.82	39.0–137.0
X*CHWGHT	X* child's weight (kg)	8,550	18.36	3.45	10.1–49.0	6,800	21.09	4.28	10.8–60.2	1,850	23.15	4.94	12.4–62.2
X*CHHGHT	X* child's height (cm)	8,600	104.58	5.39	85.0–120.0	6,800	111.81	5.50	85.0–150.0	1,850	116.60	5.45	95.5–160.0
X*CHBMI	X* child's body mass index	8,500	16.72	2.43	8.4–47.5	6,750	16.77	2.48	7.9–50.2	1,850	16.93	2.73	7.9–45.2
X*CHMUAC	X* child's middle upper arm circumference (cm)	8,550	18.12	2.09	11.0–39.0	6,800	18.83	2.31	11.0–34.1	1,850	19.33	2.37	11.0–32.5
X*CHCRFM	X* child's head circumference (cm)	850	49.63	2.31	42.0–57.1	700	50.29	2.13	42.0–62.5	250	50.44	2.00	42.0–56.8

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Where variable names and labels are provided, the data wave is not specified. For wave-specific variable names, replace the asterisk (*) with the number 3 for the preschool wave, with the number 4 for the kindergarten 2006 wave, and with the number 5 for the kindergarten 2007 wave. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table. SD = standard deviation.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Table 39. Correlations between the two values for each physical measurement by variable, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

		Preschool		Kindergarten 2006		kindergarten 2007	
Variable name	Variable label	*n*	Correlation between measurements	*n*	Correlation between measurements	*n*	Correlation between measurements
X*BIOMWT	X* biological mother's weight (kg)	7,550	1.000	5,750	1.000	1,500	1.000
X*CHWGHT	X* child's weight (kg)	8,500	0.999	6,800	1.000	1,850	1.000
X*CHHGHT	X* child's height (cm)	8,600	0.997	6,800	1.000	1,850	1.000
X*CHMUAC	X* child's middle upper arm circumference (cm)	8,500	0.980	6,750	0.997	1,850	0.998
X*CHCRFM	X* child's head circumference (cm)	850	0.978	700	0.997	250	0.996

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Where variable names and labels are provided, the data wave is not specified. Analysts looking for wave-specific variable names should replace the asterisk (*) with the number 3 for the preschool wave, with the number 4 for the kindergarten 2006 wave, and with the number 5 for the kindergarten 2007 wave. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only cases with two valid observations for each measurement and a valid weight (W*R0) are included in this table. For each wave, *n* represents the number of cases with two nonmissing values.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

5.2 Motor Skills Assessment

5.2.1 Fine Motor Skills

Fine motor skills are those that involve the small muscle groups in the body. The ECLS-B measured both fine and gross motor skills at each wave of data collection. At 9 months and 2 years, motor skills were assessed using the BSF-R. At preschool and the two kindergarten waves, a motor assessment battery was created and administered primarily from the Early Screening Inventory-Revised (ESI-R).[72] See Ketchie et al. (2003) for details about the development of the fine motor items for use in the ECLS-B. The items used in the ECLS-B are described in the following sections.

5.2.1.1 Building Block Items and Their Scores

The preschool fine motor assessment included two items involving the use of blocks to build structures drawn from the ESI-R preschool version: "build a tower" and "build a gate." Of these, the kindergarten assessments included only the build-a-gate item. A description of these items and their scoring is provided in table 40. For both of these items, the child was presented with a set of wooden blocks. For the build-a-tower item, the child was instructed to build a tower with a set of 10 blocks. The child received 1 point for each block correctly positioned to make a tower (up to 10 blocks, or points). The child was allowed up to two trials, with the higher score being retained. For the build-a-gate item, the child was presented with a set of five wooden blocks. Then, with another set of blocks, the FI demonstrated how to build a gate following the diagram in the Child Assessment Booklet. When the model was completed, it was left assembled, and the FI asked the child to "make one just like mine." The child received a passing score if his or her structure looked like the gate the FI built. Children were allowed only one trial for the gate item but could work on building the gate until they were satisfied with what they had produced.

Table 40. ECLS-B fine motor items using blocks, by item and description: 2005–06, 2006–07, and 2007–08

Item	Description of scoring	Scoring
Build a tower	The number of blocks (up to 10) that the child successfully uses in building the tower.	0–10 (continuous)
Build a gate	Child receives credit for a pass if the child's gate looks like the model provided by the field interviewer.	0 = fail; 1 = pass

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

The sample size and percentage of children passing the build-a-gate item during the preschool, kindergarten 2006, and kindergarten 2007 waves are included in table 41.

[72] The ESI-R consists of two separate versions, the Early Screening Inventory-Preschool (ESI-P), for children ages 3 to 4½, and the Early Screening Inventory-Kindergarten (ESI-K), for children ages 4½ to 6. There is overlap between these two forms. For ease of reference, we refer to the overall instrument as the ESI-R, indicating the form (preschool or kindergarten) within the text.

Table 41. Summary statistics for the build-a-gate fine motor item, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

		Preschool		Kindergarten 2006		Kindergarten 2007	
Variable name	Item	*n*	Percent	*n*	Percent	*n*	Percent
C3FM2GTE/ C*FM1GTE	Build a gate, percentage who passed	8,400	48.3	6,800	78.1	1,900	87.5

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. For wave-specific variable names for C*FM1GTE, replace the asterisk (*) with the number 4 for the kindergarten 2006 wave and with the number 5 for the kindergarten 2007 wave. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

The weighted mean scores, standard deviations, and ranges for the build-a-tower item, as well as a composite score that provides a summary of the child's performance on the building block items during the preschool wave of data collection (X3FMBLCK), are shown in table 42. To create this composite, the build-a-tower item was recoded to be pass/fail. To "pass," a child needed to construct a tower of 10 blocks. A tower of nine blocks or fewer, or no tower at all, was coded as "fail." The composite is a sum of the recoded build-a-tower and build-a-gate scores. The block-building composite score ranges from 0 to 2, with 0 points assigned if the child failed both tasks, 1 point assigned if the child successfully completed only one of these tasks, and 2 points assigned if the child passed both tasks.

Table 42. Fine motor items using blocks, by item, preschool wave: 2005–06

Item	*n*	Mean	*SD*	Range
Build a tower (C3FM1TWR)	8,600	9.13	2.06	0–10
Overall block building items (X3FMBLCK)	8,350	1.26	0.69	0–2

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data. Only those cases with a valid weight are included in the table. *SD* = standard deviation.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

5.2.1.2 Copy Forms and Their Scores

The child fine motor assessment also included a series of items in which the child was shown a drawing and asked to replicate the drawing in pencil on a blank page of the Child Assessment Booklet. The specific items and their scoring are shown in table 43. On the preschool assessment, the child was provided with seven specific forms to draw from the ESI-R: vertical line, horizontal line, circle, square, triangle, asterisk, and cross (or addition sign). The kindergarten assessments included the square, triangle, and asterisk items, as well as a circle-square item from the ECLS-K. As mentioned above, the child drew these forms directly in the Child Assessment Booklet. The drawings were then scored centrally at RTI. Only one trial was required for each item, but children could make as many attempts as they wished. In the case of multiple attempts, the FI numbered the attempts in the Child Assessment Booklet. If multiple

attempts were made and labeled, the second attempt was the one scored. During the preschool wave, if there were multiple attempts but they were not labeled, the item was not scored. During the kindergarten waves, if there were multiple attempts that were not labeled, but all attempts would receive the same score, a score was given. When there were multiple attempts without labels and each attempt would receive different scores, the item would be scored as "uncodeable." Each of these items was scored as a pass or a fail.

Table 43. ECLS-B copy forms item variable names and scoring, by item, for the ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Item	Preschool wave	Kindergarten 2006 and 2007 waves	Description of scoring	Scoring
Vertical line	C3FM3VLN	†	Scored as pass or fail	0–1
Horizontal line	C3FM4HLN	†	Scored as pass or fail	0–1
Circle	C3FM5CRL	†	Scored as pass or fail	0–1
Addition sign/cross	C3FM6ADD	†	Scored as pass or fail	0–1
Square	C3FM7SQR	C*FM2SQR	Scored as pass or fail	0–1
Triangle	C3FM8TRI	C*FM3TRI	Scored as pass or fail	0–1
Asterisk	C3FM9AST	C*FM4AST	Scored as pass or fail	0–1
Circle-square	†	C*FM5CSQ	Scored as pass or fail	0–1
Curved path	C3FM10PA	†	Scored as fail, partial, or full pass	0–2

† Not applicable; measure not administered in this wave.

NOTE: Where variable names include an asterisk (*), the data wave is not specified. For wave-specific variable names, replace the asterisk with the number 4 for the kindergarten 2006 wave and with the number 5 for the kindergarten 2007 wave.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

In addition, during the preschool wave, children were asked to draw a line following a curved path (the "curved path" item from the Bruininks-Oseretsky Test). All children were allowed a single trial, but a second trial was granted to children who specifically requested an opportunity to "do it over." In cases where children completed two trials, the second trial was scored. The item was scored as a full or partial pass (2 or 1 point) based on how often the child's line was within one-half inch of the path borders—full credit was awarded if the line followed within the path (i.e., did not fall more than ½ inch outside of the path) and partial credit was awarded if the line fell outside of the path once or twice). If a line fell outside of the path more than once or twice, the child earned 0 points.

Each of the fine motor drawing items was scored by specially trained coders using a protocol adapted from the appropriate source protocol (e.g., ESI-R, ECLS-K, or Bruininks-Oseretsky). The coders were evaluated for reliability on a regular basis against standard coder scores (see chapter 8). Items also may have been scored as not codeable (and set to missing in the data) if more than one attempt was made but the attempts were not numbered and each attempt would have received a different score. (If the attempts were not numbered but the score of each attempt was identical, that score was used as a valid score for the item.) Additionally, for each figure attempted, the FI recorded in the Child Assessment Booklet which hand the child used.

Summary statistics for each of the copy forms items are shown in table 44. In addition, each child's performance across all of the copy forms items (seven during the preschool wave,[73] four during the kindergarten 2006 and kindergarten 2007 waves) was summarized by summing the number of items the child received credit for or passed. Summary statistics for this item (X*FMFORM) are included in

[73] The preschool curved path item was not included in the overall preschool copy forms fine motor score; only dichotomous items are included in the overall score.

table 45.

Table 44. Summary statistics for copy forms fine motor items, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

		Preschool		Kindergarten 2006		Kindergarten 2007	
Variable name	Item	*n*	Percent	*n*	Percent	*n*	Percent
C3FM3VLN	Vertical line, percentage who passed	8,250	83.5	†	†	†	†
C3FM4HLN	Horizontal line, percentage who passed	8,250	81.4	†	†	†	†
C3FM5CRL	Circle, percentage who passed	8,500	73.4	†	†	†	†
C3FM6ADD	Addition sign/cross, percentage who passed	8,500	39.3	†	†	†	†
C3FM7SQR/ C*FM2SQR	Square, percentage who passed	8,400	20.8	6,750	63.5	1,850	69.1
C3FM8TRI/ C*FM3TRI	Triangle, percentage who passed	8,400	15.0	6,800	62.5	1,850	74.6
C3FM9AST/ C*FM4AST	Asterisk, percentage who passed	8,400	16.2	6,800	69.5	1,850	80.7
C*FM5CSQ	Circle-square, percentage who passed	†	†	6,800	34.8	1,850	48.9
C3FM10PA	Curved-path, percentage receiving full credit	8,450	57.6	†	†	†	†

† Not applicable; measure not administered in this wave.

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Where variable names include an asterisk (*), the data wave is not specified. For wave-specific variable names, replace the asterisk with the number 4 for the kindergarten 2006 wave and with the number 5 for the kindergarten 2007 wave. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Table 45. Summary statistics for overall copy forms fine motor score, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Variable name	Item	*n*	Mean	*SD*	Range
X3FMFORM	Overall copy forms	7,950	3.33	1.55	0–7
X4FMFORM	Overall copy forms	6,700	2.31	1.33	0–4
X5FMFORM	Overall copy forms	1,850	2.73	1.14	0–4

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table. *SD* = standard deviation.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

5.2.2 Gross Motor Skills

Gross motor skills are those that involve the larger muscle groups of the body, such as legs and arms. During the preschool and kindergarten child assessments, FIs assessed a number of gross motor skills. Children's performance on these items was recorded in the Child Assessment Booklet, prior to entry into the Child CAPI Instrument; scores were entered into the Child CAPI Instrument before the FI left the home. The following paragraphs offer a brief description of each gross motor item. It is important to note that while effort was made to include all children to the greatest extent possible, some children with disabilities were not administered these items, either because of their physical limitations (e.g., they used a wheelchair) or because of safety concerns (e.g., a child had a severe visual impairment).

5.2.2.1 Jumping

This item was drawn from the Bruininks-Oseretsky Test of Motor Proficiency (Bruininks 1978). For this item, the child was asked to jump from a standing position as far as he or she could. The FI placed a 6-foot length of tape on the floor, with perpendicular pieces of tape on either end. The FI modeled the jump for the child to be sure that the child bent his or her knees and swung his or her arms appropriately and then asked the child to jump straight off the starting line (one of the tape lines perpendicular to the 6-foot length of tape). The child was given two trials. On each trial, the distance between the starting line and the part of the child's body that landed closest to the starting line (e.g., the closest foot, or hand if the child landed and fell backward) was measured and recorded. The longest distance was entered into CAPI by the FI prior to leaving the home.

5.2.2.2 Balancing on One Foot

This item was drawn from the Early Screening Inventory-Kindergarten (ESI-K). For this item, the child was asked to balance on one foot for as long as possible, up to 10 seconds. The FI modeled the appropriate posture for the child and then asked the child to balance. The child was given up to three trials for each foot, although additional trials were not required once the child balanced on that foot for 10 seconds. The FI recorded in the Child Assessment Booklet which

foot the child chose to balance on first. The child's performance on each trial attempted was recorded and entered into CAPI. The data file provides three variables for each foot indicating whether the child balanced successfully on it for 10 seconds, the maximum amount of time the child balanced and, for those children who balanced successfully, the number of trials taken. The data file also provides a variable indicating on which foot the child balanced first.

5.2.2.3 Hopping

This item also was taken from the ESI-K. The child was asked to hop five times on each foot without allowing the second foot to touch the floor. The FI modeled the task for the child. The child was then given up to three trials for each foot to achieve five hops. The child was not required to complete any more trials for a given foot once he or she had successfully hopped on it five times without allowing the other foot to touch the ground. As with the balancing task, the FI recorded in the Child Assessment Booklet which foot the child chose to hop on first. The child's performance on each trial attempted was recorded and entered into CAPI. The data file provides three variables for each foot indicating whether the child successfully hopped five times on it, the maximum number of hops the child had and, for those children who hopped five times successfully, the number of trials taken. The data file also provides a variable indicating on which foot the child first hopped.

5.2.2.4 Skipping

This item was drawn from the ESI-K. The child was asked to skip and was scored for skipping up to eight consecutive skips. The FI modeled the task for the child to ensure that the child was skipping rather than galloping or using some other motion. The child was given only one trial on this item, but if he or she stopped prior to demonstrating eight consecutive skips, the FI asked the child to "keep going." The FI then recorded whether the child successfully completed eight consecutive skips. The data file includes the child's score as a pass or fail.

5.2.2.5 Walking Backward

This item was taken from the ECLS-K. For this item, the child was asked to walk backward along a line on the floor and was scored for up to six steps. To administer this item, the FI used the 6-foot length of tape that was placed on the floor for administering the jump item. The FI then modeled the task for the child, pointing out that the FI's feet were staying on the line. The child was then given one trial for the item and again told to "keep going" if he or she stopped walking before taking six steps. The FI scored the trial on a 4-point scale, with a score of 0 indicating that the item was attempted but the child was unable to complete any steps on the line, a score of 1 indicating that the child walked backward with more than two steps off the line, a score of 2 indicating that the child walked backward with one or two steps off the line, and a score of 3 if the child was able to walk backward six steps with both feet on the line for all steps.

5.2.2.6 Bean Bag Catch

The final gross motor skills item administered in the home visit was the bean bag catch (taken from the Movement Assessment Battery for Children; Henderson and Sugden 1992). To administer this item, the FI used the 6-foot length of tape that was placed on the floor for administering the jump item. The FI stood at one end of the line and asked the child to stand on the other end. The child was told to catch the bean bag with two hands but to stay behind the line. The FI then gave two practice trials and provided feedback as necessary (e.g., reminded the child to use two hands or to stay behind the line). The child was then given five trials, and the number of catches across the five trials was recorded. The FI was instructed to repeat, and not count, a trial if the toss from the FI was above the child's shoulders, below the child's waist, or out of the child's reach.

5.2.2.7 Gross Motor Scores

The gross motor items included some tasks for which the child was given multiple trials and some for which the child was given only a single trial. For items allowing single trials, the child's score was based on the completed trial (e.g., skip, walk backward, bean bag catch). Two items (balancing on one foot and hopping on one foot) allowed the child to attempt multiple trials; the child's performance on each trial was recorded and contributed to how the item was scored. For these items, the FI recorded which foot was attempted and how long the child balanced (for the balance item) or how many hops were completed (for the hopping item) for each trial, up to three trials per foot on each item. The data file includes a number of variables that describe the child's performance on the gross motor items, as shown in table 46. The pass/fail items on the data file are presented as 0 = fail and 1 = pass. Also, in addition to the ECLS-B reserve codes (-9 for not ascertained and -7 for refused) for each item, additional codes are possible for nonscored responses (96 = Not administered, 97 = No response, and 99 = Physical limitation). Summary statistics for the gross motor items are presented in tables 47 and 48.

Table 46. Variable name, label, description, and scoring for gross motor items in the ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Variable name	Variable label	Description	Values
Jump item			
C*GM1SCR	C* gross motor jump	Distance the child jumped from a standing start, measured in inches. Child had two trials. Farthest distance recorded.	0–50 inches
Balance items			
C*GM2RFT	C* gross motor balance score (right foot)	Child balanced on right foot for 10 seconds on at least one of the three trials.	1 = pass; 0 = fail
C*GM2RTR	C* gross motor balance right foot trial # success	Number of trials taken for the child to successfully balance for up to 10 seconds on the right foot.	1 = 1 trial 2 = 2 trials 3 = 3 trials
C*GM2RSE	C* gross motor balance max count right foot	Maximum count across all trials for right foot.	0–10
C*GM2LFT	C* gross motor balance score (left foot)	Child balanced on left foot for 10 seconds on at least one of the three trials.	1 = pass; 0 = fail
C*GM2LTR	C* gross motor balance left foot trial # success	Number of trials taken for the child to successfully balance for up to 10 seconds on the left foot.	1 = 1 trial 2 = 2 trials 3 = 3 trials
C*GM2LSE	C* gross motor balance max count left foot	Maximum count across all trials for left foot.	0–10
C*GM2WFT	C* gross motor balance which foot	Which foot child started with for balancing.	1 = right; 2 = left
Hop items			
C*GM3RFT	C* gross motor hop score (right foot)	Child hopped on right foot five times on at least one of the three trials.	1 = pass; 0 = fail
C*GM3RTR	C* gross motor hop right foot trial # success	Number of trials taken for the child to successfully hop up to five times on the right foot.	1 = 1 trial 2 = 2 trials 3 = 3 trials
C*GM3RTM	C* gross motor hop max count right foot	Maximum count across all trials for right foot.	0–5
C*GM3LFT	C* gross motor hop score (left foot)	Child hopped on left foot five times on at least one of the three trials.	1 = pass; 0 = fail
C*GM3LTR	C* gross motor hop left foot trial # success	Number of trials taken for the child to successfully hop up to five times on the left foot.	1 = 1 trial 2 = 2 trials 3 = 3 trials
C*GM3LTM	C* gross motor hop max count left foot	Maximum count across all trials for left foot.	0–5
C*GM3WFT	C* gross motor hop which foot	Which foot child started with for hopping.	1 = right; 2 = left

See notes at end of table.

Table 46. Variable name, label, description, and scoring for gross motor items in the ECLS- B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08—Continued

Variable name	Variable label	Description	Values
Skip item			
C*GM4SCR	C* gross motor skip	Child was asked to skip for at least eight consecutive steps. Child received one trial. Trial was scored as pass or fail.	1 = pass; 0 = fail
Walk backward item			
C*GM5SCR	C* gross motor walk backward	Child walked backward along a line for at least six steps. Child received one trial.	0 = attempted, failed to walk backward at all 1 = failed, walked backward with more than two steps off line 2 = failed, walked backward with one or two steps off line 3 = pass, walked backward six steps with feet on line
Bean bag catch item			
C*GM6SCR	C* gross motor bean bag catch	Child was tossed a bean bag up to five times for a single trial. Scored as the number of times the bag was caught.	0–5

NOTE: Where variable names and labels are provided, the data wave is not specified. For wave-specific variable names, replace the asterisk (*) with the number 3 for the preschool wave, with the number 4 for the kindergarten 2006 wave, and with the number 5 for the kindergarten 2007 wave.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Table 47. Summary statistics for gross motor items jump and bean bag catch, ECLS- B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Variable name	Variable label	Preschool				Kindergarten 2006				Kindergarten 2007			
		n	Mean	*SD*	Range	*n*	Mean	*SD*	Range	*n*	Mean	*SD*	Range
C*GM1SCR	C* gross motor jump	8,350	25.37	8.85	0–49.9	6,750	30.66	8.95	0–60.0	1,850	33.47	8.96	0–67.2
C*GM6SCR	C* gross motor bean bag catch	8,450	3.68	1.38	0–5.0	6,800	4.08	1.18	0–5.0	1,850	4.32	1.08	0–5.0

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Where variable names and labels are provided, the data wave is not specified. For wave-specific variable names, replace the asterisk (*) with the number 3 for the preschool wave, with the number 4 for the kindergarten 2006 wave, and with the number 5 for the kindergarten 2007 wave. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table. *SD* = standard deviation.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Table 48. Summary statistics (percent passing) for gross motor items balance, hop, skip, and walk backward, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Variable name	Variable label	Preschool		Kindergarten 2006		Kindergarten 2007	
		n	Percent	*n*	Percent	*n*	Percent
C*GM2RFT	C* gross motor balance, right foot, percent passed	8,200	46.7	6,750	76.8	1,850	83.1
C*GM2LFT	C* gross motor balance, left foot, percent passed	8,150	44.5	6,750	75.8	1,850	82.1
C*GM3RFT	C* gross motor hop, right foot, percent passed	7,900	65.1	6,700	88.4	1,850	95.4
C*GM3LFT	C* gross motor hop, left foot, percent passed	7,850	60.3	6,700	85.9	1,850	93.3
C*GM4SCR	C* gross motor skip, percent passed	7,550	25.1	6,500	48.3	1,800	62.3
C*GM5SCR	C* gross motor walk backward, percent walking backward six steps with feet on line	8,100	34.7	6,700	41.7	1,850	51.5

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Where variable names and labels are provided, the data wave is not specified. For wave-specific variable names, replace the asterisk (*) with the number 3 for the preschool wave, with the number 4 for the kindergarten 2006 wave, and with the number 5 for the kindergarten 2007 wave. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Chapter 6
The Two Bags Task and the Reading Aloud Profile–Together Coding in the Preschool Wave of the ECLS-B

This chapter begins with an overview of the Two Bags Task and how it was implemented during the preschool wave of the ECLS-B, including a summary of scoring procedures, a description of how the task was administered in the field, and an account of the training provided for field interviewers and coding staff. Following the overview of the Two Bags Task is a description of the quality control procedures employed and a summary of how coder reliability was maintained. Next, a description of the Two Bags scales is provided, including correlations among scales and between the scales and composite scores. This is followed by a comparison of the implementation and scoring of the Two Bags Task at the 2-year and preschool waves. The chapter concludes with a description of the Reading Aloud Profile–Together (RAPT) coding implemented for a subsample of the preschool wave Two Bags Task data.

6.1 Two Bags Task

Parent-child interaction is a key aspect of children's socioemotional development and contributes to children's early learning experiences. Consequently, it was important to include a direct measure of children's socioemotional development in the ECLS-B. During the preschool data collection, parent-child interactions were assessed using the same method employed during the 2-year data collection: the Two Bags Task. This task is a modification of the Three Bags Task, which has been used with success in other large-scale studies, including the Early Head Start (EHS) Research and Evaluation, and the National Institute of Child Health and Human Development (NICHD) Early Child Care Study.

The Three Bags Task is a semistructured interaction between the parent and child in which the dyad is given 15 minutes to play with the objects found in three separate numbered bags. Because the ECLS-B protocol was such that field interviewers would be spending a long time in study families' homes to complete many of the study components, it was decided to modify the Three Bags Task to two bags, making certain to retain the joint book reading activity and to decrease the target duration of interaction. Thus, the Two Bags Task consisted of two bags/activities and took about 10 minutes to complete. In the preschool wave of the ECLS-B, bag number 1 contained the book *Corduroy*, by Don Freeman (1968), and bag number 2 contained Play-Doh®, a rolling pin, and cookie cutters. The dyad was told that they had 10 minutes to play with the two bags, the only restriction being that they had to play with the contents of the bags in numerical order. The parent and child were video recorded while they

engaged in the activities. The recordings (on DVD) were sent to RTI, where the parent-child interaction was scored by trained coders.

6.2 Two Bags Task Coding in the Preschool Data Collection

The Two Bags Task rating scales include five parent rating scales and three child rating scales. Each scale is a 7-point Likert-type rating scale that ranges from very low (1) to very high (7). The DVDs were scored using rating scales adopted, with permission, from those developed for the EHS study by Fauth, Brady-Smith, and Brooks-Gunn (2003). The EHS study was a modification of the EHS 14-month Child-Parent Interaction Rating Scales for the Three Bags Task assessment developed by Ware et al. (1998) and the manual for scoring free play developed by Brooks-Gunn et al. (1992). Each rating level was well described in the coding manual with specific examples to illustrate the concept and target behaviors.

To score a DVD (i.e., a Two Bags interaction), the coder watched the interaction and observed the target behaviors measured by each scale, making notes that would help with the coding for each scale. When the DVD was finished, the coder assigned a score for each scale based on the observations and notes made while watching the DVD. Coders were trained to provide a score for each scale independently of the others (i.e., a high score on a positive scale did not necessarily mean a low score on a negative scale). This also allowed for coders to provide scores on some scales but not others if a score could not be determined or justified.[74]

The five parent rating scales are as follows:

Parental Emotional Supportiveness: This scale focuses on the parent's emotional availability and physical and affective presence during the task. Emotionally supportive parenting involves (1) providing a secure base from which the child can explore, and (2) displaying emotional support and enthusiasm for the child and his or her autonomous work.

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

[74] In reviewing the DVDs, in some instances a brief interruption in sound or light quality, poor image of the child's or parent's face, or other factors may have limited a coder's ability to provide an accurate score for a given scale or scales where the score would depend on or be greatly influenced by what the coder could not see or hear. This happened rarely (fewer than 50 cases out of approximately 7,650 with at least one scale scored), however, because the scales do not rely on a single indicator; thus, it is more likely that poor sound or video quality would have affected the coder's ability to provide a score on only a subset of the scales.

Parental Stimulation of Cognitive Development: This scale focuses on the parent's effortful teaching to enhance perceptual, cognitive, and language development. A stimulating parent is aware of the child's developmental level and aims to bring the child to the next level. If the focus or method of stimulation is not matched to or slightly above the child's developmental level or interest, then the parent's behavior is not seen as stimulating cognitive development. In the Two Bags Task, the environment is somewhat constrained (i.e., the dyad uses materials specifically provided for the task), so there is a reduced likelihood that the parent would try to stimulate the child using something inappropriate (i.e., any item not included in the bag); however, because data collection occurred in the home, it was possible for the parent to introduce an object that was not intended to be included in the task.

Parental Negative Regard: This scale reflects the parent's expression of discontent with, anger toward, disapproval of, or rejection of the child. The key is to score parental negative regard from the point of view of the child; it was scored independently of the parent's positive behaviors captured in the emotional supportiveness scale. That is, this scale focuses on the parent's negative behavior only—it is not scored as low emotional supportiveness.[75]

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

The three child rating scales are as follows:

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

Child Negativity Toward Parent: This scale measures the degree to which the child shows anger, hostility, or dislike toward the parent. At the high end, the child is repeatedly and overtly angry with the parent. The important point is that at this age,

[75] A parent could, in the course of one 10-minute interaction, display a high amount of emotional supportiveness and a high amount of negative regard for the child (e.g., a passionate and moody parent). Likewise, a parent could display little emotional supportiveness and also little negative regard (e.g., a neutral or emotionally reserved parent). Therefore, while most parents were probably high on one measure and low on the other, these scales did, in fact, function independently.

the child may express negativity toward the parent by hitting an object, the floor, or himself or herself, by pushing the parent away, by throwing a toy, or by using a negative expression to communicate that he or she wants or does not want something (e.g., "No!"). Therefore, the context of the negative expression should be taken into account when determining the extent to which it is directed toward the parent.

Child Quality of Play: This scale assesses the child's sustained involvement with objects and the quality of his or her play. Quality of play encompasses three components: attention to play objects, self-direction, and complexity of play. These three aspects of play quality were coded both during the book reading task and the play task.

6.2.1 Two Bags Task Protocol for In-Home Administration

The Two Bags Task is a standardized, semistructured play task administered during the home visit and recorded for later scoring by trained coders. The Two Bags Task administration during the home visit was standardized to ensure that all interviewers administered the task in the same way to all parent-child pairs. To ensure this standardization, step-by-step Two Bags Task administration instructions were given to field interviewers. These instructions included a verbatim script that was read to the parent. Interviewers also asked parents whether they had previously read *Corduroy* to their child, and if so, how often. Interviewers were expected to record parents' answers in check boxes on the administration pages in the Child Assessment Booklet, making sure to record verbatim answers related to frequency.[76]

In the case of twins, the interviewer administered the Two Bags Task separately for each twin and recorded the interaction for each twin on separate DVDs. This created the problem of familiarity with the storybook as a confounding variable for the second twin. It was possible that on the second reading of the storybook, the parent would alter the reading in some systematic way, but it was also possible that the parent's response to the task may be altered by fatigue. Therefore, the administration of the task was counterbalanced with field interviewers instructed to administer the Two Bags Task first to the first-born twin on odd-numbered days and first to the second-born twin on even-numbered days. As a result, the effect of familiarity was spread across twin pairs such that half of the older twins would receive the task first and half of the younger twins would receive the task first. Order of administration can thus be used as a covariate to control for administration order. Field staff recorded in the Child Assessment Booklet which twin had been administered the Two Bags Task first and this is included in the data set.

After completion of the home visit, the field interviewer sent the Two Bags Task DVD and the Child Assessment Booklet, along with other data collection materials, to RTI's home

[76] Items about frequency of reading the book used in the Two Bags Task were added after data collection was begun. As a result, data on the frequency of reading the book were not obtained for approximately 800 children. A small number of cases (fewer than 50) included responses that could not be coded due to vagueness (e.g., "lots of time"). The remaining cases with missing data include refusals or other nonresponse.

office for receipting and scoring by trained coders. Once scoring was completed on a case, the coder entered the data into the coding data file. Accuracy of data entry was checked in conjunction with an ongoing reliability check conducted by a standard coder for approximately 5 percent of each coder's cases.

6.2.2 Two Bags Task Field Staff Training, Coding Trainer Training, and Coder Training

Field interviewers were responsible for the administration and taping of the Two Bags Task. However, coding of the Two Bags Task interaction was completed by staff located at RTI. The training for coding staff followed a "train the trainer" model, with key project staff, including a coding task leader and the initial standard coders,[77] receiving training on scoring directly from the originators of the system at Columbia University Teachers College (CUTC). Once certified by the CUTC trainer, the aforementioned standard coders then led the training of the field coders.

6.2.2.1 Field Staff Training

Because the Two Bags Task has a standardized protocol, field interviewers were trained to administer the task in a consistent way and to prepare a high-quality recording that could be accurately scored back at the home office. Training for the administration of the Two Bags Task was part of the week-long training that field interviewers attended to prepare for data collection. The Two Bags Task was recorded in the respondent's home, so field interviewers were trained to use a tripod and digital video recorder to create DVDs that could be viewed later for scoring in the home office. They were also trained to create a voice stamp on each DVD so that they could (1) ensure the equipment was working properly with each child and (2) properly identify the case.

6.2.2.2 Coding Trainer Training

The initial standard coders and the coding task leader attended the trainer training conducted by Dr. Rebecca Fauth in October 2004. Dr. Fauth, a consultant who provided training and expertise to the EHS Research and Evaluation coding team, trained the ECLS-B standard coders using the procedures developed for the EHS Research and Evaluation (Fauth, Brady-

[77] The coding of Two Bags Task interactions during the preschool wave used a standard-coder model, in which a small number of standard coders double-coded a percentage of the work completed by field coders. In this model, the standard coders were assumed to be the gold standard in application of the coding scales used. Because of the number of cases to be coded and the size of the coding staff, more than one standard coder was used throughout the duration of coding operations. Initially, three standard coders were certified. One standard coder left RTI shortly after coding began, and the cases completed by this standard coder were recoded by another standard coder. A second standard coder left RTI after 90 percent of the cases were coded, and approximately 25 percent of reliability coding was completed. Subsequently, two new standard coders were trained and certified as standard coders to complete the reliability coding of cases. A total of 27 field coders and 5 standard coders (no more than 3 serving at any one time) were trained and certified in their roles, although staff attrition resulted in a final set of 26 field coders and 4 standard coders contributing scores for the task.

Smith, and Brooks-Gunn 2003).[78] This 3-day training session took place at RTI, with key members of the ECLS-B research team (the two initial standard coders;[79] the coding task leader, who was also a standard coder; and the project director) receiving training. Training included an explanation of each scale and the meaning of various scores on that scale, viewing of exemplar tapes from the EHS project, individual coding practice, and group discussion. Training emphasized that the coders must provide specific examples for the scores they assigned to each scale (written in complete sentences) and be able to justify their decisions.

It should be mentioned that the procedure for the Two Bags Task as it was used in the EHS project differed from the procedure used in the ECLS-B; the task as used in the EHS study did not include a book and so did not include the requirement that the parent and child begin the task with the book, as was done in the ECLS-B. These procedural differences were discussed by the ECLS-B staff and the CUTC trainer. The coding scales focus on the same constructs in both studies; to accommodate the difference in procedures, the RTI team worked with the CUTC trainer to develop additional scoring points for each scale that would be consistent with the approach taken in the EHS study and appropriate to the ECLS-B implementation. In this way, ECLS-B coders learned the EHS coding scheme as it applied to both the EHS Three Bags interactions and the ECLS-B Two Bags interactions.

At the completion of this training, the standard coders were certified by Dr. Fauth. To be certified, each coder had to score a set of five interactions that had already been scored by EHS coders. Each coder's score was compared to the EHS score on each scale (40 scores total, 5 cases with 8 scales each). Coder scores were considered to be in agreement if they were within 1 point of the EHS-established score. The percent agreement for each coder was calculated as the percentage of items (from the 40 scores) for which the two scores were within 1 point of each other. To be certified, coders had to demonstrate 85 percent agreement with the EHS coders on all scales.

Due to attrition among the standard coders (only one of the original three remained on staff; the coding task leader had also left the project), a second set of standard coders was trained and certified in September 2006 by the original coding task leader (who had been trained and certified by Dr. Fauth in 2004).[80] These new standard coders were recruited from the initial set of field coders (whose training is described below), based on the strength of their reliability as assessed during field coding. This training followed the agenda and approach used during the initial training conducted by Dr. Fauth. The additional standard coders were certified using a set of 10 EHS cases and the accompanying scale scores from the initial training (including 5 that were used for initial standard coder certification), with a criterion of 90 percent agreement

[78] The initial training by Dr. Fauth was timed to occur prior to the preschool field test so that coding of field test cases could be completed prior to the national preschool data collection.

[79] There were three standard coders initially trained (including the coding task leader). One of the standard coders left the project shortly after coding began. Subsequently, cases for which she provided the standard scores were coded by a remaining standard coder, and that set of scores was used to check the field coders' reliability and was maintained on the data file.

[80] At no time were more than three standard coders active on the project.

(within 1 point) on each scale.[81] Certification results for the four standard coders who contributed data to the final data set are provided, by scale, in table 49.

Table 49. Percentage agreement (within 1 point) for all Two Bags Task standard coders against established EHS case scores, by scale, preschool data collection: 2005–06

Scale	*n*	Average percent agreement	Range of percent agreement
Parental emotional supportiveness	4	100	100
Parental stimulation of cognitive development	4	100	100
Parental intrusiveness	4	90	80–100
Parental negative regard	4	100	100
Parental detachment	4	100	100
Child engagement	4	100	100
Child quality of play	4	100	100
Child negativity	4	100	100

NOTE: Data are for the four standard coders who contributed data in the final data file. EHS = Early Head Start.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

6.2.2.3 Coder Training

The training for 17 field coders took place in September 2005 and was led by the coding task leader, with assistance from the principal investigator and standard coders.[82] The coders hired and trained to complete coding for the Two Bags Task were also trained to complete coding on the PreLAS Let's Tell Stories items and the fine motor items (see chapter 8 for a discussion of the training and reliability procedures for those activities). During a 3-day training session, the coding task leader followed the same training agenda used during the initial training for the standard coders. She provided an overview of the coding task, reviewed the rating scales, and provided individual practice and group discussions. This training was informed by the standard coders' experience in scoring cases coded as part of the preschool field test. After training, each field coder was assigned certification cases—five cases drawn from incoming ECLS-B preschool cases that had been coded by the three previously trained and certified standard coders. Each field coder independently scored each of the five cases on each scale. The coders' scores were then compared to the scores established by the three standard coders. To be certified, each field coder was required to achieve 80 percent agreement (within 1 point) or

[81] Note that this criterion is higher than was used for initial certification. One reason for this is to further tighten the initial reliability of these coders given their role as the standard against which field coders were held. The second reason is practical; a set of 10 cases was used for certification purposes, so 90 percent agreement on each scale required agreement (within 1 point) on at least 9 of these 10 cases.
[82] One of these field coders left the project shortly after coding began. All of this coder's cases were subsequently recoded by other field coders.

higher with the established scores across all scales and cases.[83] The percent agreement for the initial field coders is shown in table 50. Note that table 50 provides percent agreement data for the certification cases for the 13 initial field coders (out of the 17 trained) who contributed scores to the final data set (as described below). As shown, the average reliability criterion of 80 percent agreement was met on each scale, across all coders, although coder performance on the scales varied (90.8 through 98.5 percent agreement). The range of agreement underscores the variability in the definitions of each scale and the degree to which the scale identifies behaviors that are generally variable. When scale definitions include fewer concrete descriptors (e.g., parental emotional responsiveness), the scales tend to be more difficult to code and, as a result, coders may demonstrate lower reliability in their coding. Additionally, when scales reflect behavior that is generally of limited variability (e.g., child negativity), coding tends to be easier, but scales reflecting behavior that can be more variable (e.g., parental emotional supportiveness, child quality of play) make agreement more challenging.

Table 50. Percent agreement (within 1 point) for Two Bags Task initial set of field coders, by scale, ECLS-B preschool data collection: 2005–06

Scale	*n*	Average percent agreement	Range of percent agreement for passing scores
Parental emotional supportiveness	13	90.8	60–100
Parental stimulation of cognitive development	13	90.8	80–100
Parental intrusiveness	13	93.9	80–100
Parental negative regard	13	95.4	80–100
Parental detachment	13	96.9	80–100
Child engagement	13	90.8	80–100
Child quality of play	13	90.8	80–100
Child negativity	13	98.5	80–100

NOTE: Data are from the 13 field coders who contributed scores to the final data set. Note that percent agreement for individual scales could be below 80 percent and the coder would still have been certified if the average across scales was greater than 80 percent; this occurred for the parental emotional supportiveness scale.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

6.2.3 Two Bags Coding Quality Control Procedures and Reliability

Each set of child assessment case materials was receipted and checked for completeness upon arrival at RTI. In the case of the Two Bags Task, this check included confirmation that a DVD was included in the package of materials and that it was appropriately labeled.

When field coders (working in a central secure location) began to work on a specific case, they noted any problems with the case, including failure to receive a DVD (i.e., there was

[83] The reliability criterion for the field coders was set lower than the reliability criteria for the standard coders. At the beginning of preschool coding operations, the intended overall reliability criterion was set at 80 percent; the criterion for standard coders (as the comparison against which field coder's reliability would be measured) therefore was actually set higher than the expected level of reliability over the course of coding operations as a means of increasing the likelihood that coders would meet or exceed the 80 percent agreement target.

an empty jewel case or no DVD accompanying the assessment materials), any technical problems with the DVD, any procedural violations in the completion of the Two Bags Task, the language used by parent and child during the interaction, and the duration of the interaction recorded on the DVD. In some cases, this review concluded that a case could not be coded because of technical problems with the recording (4.9 percent of cases with DVDs) or because the language spoken was neither English nor Spanish (1.3 percent of cases with DVDs). No cases were excluded due to short duration (the interaction was less than 5 minutes in duration). If the case was able to be coded, the field coder proceeded with coding the case as described below.

6.2.3.1 General Reliability Procedures

Throughout the majority of the Two Bags Task coding (September 2005 through April 2006), 5 percent of each coder's completed cases were assigned to a standard coder for double-coding—that is, independent coding of the cases without knowledge of the field coder's scores. The two scores from this 5 percent sampling of each coder's work were then compared, and the percent agreement for each coder, on each scale, was calculated. In this comparison, agreement was defined as two scores (field coder and standard coder) within 1 point. Each coder's percent agreement was calculated for each of the 8 scales as the number of cases with field coder scores within 1 point of the standard coder scores divided by the total number of cases in the 5 percent sample. When a coder's reliability for a given scale was below 80 percent, the standard coder reviewed with the field coder those cases for which the field coder's score differed from the standard coder's score by more than a single point. In addition, for field coders whose percent agreement was less than 80 percent on a scale (or scales), the next three to four cases completed by the field coder were reviewed by the standard coder until agreement on each scale was demonstrated by the field coder (against the standard coder's scores).[84] In all cases, irrespective of percent agreement score, the standard coder's scores were assigned to the double-coded case, replacing those of the field coder even when the percent agreement between scores was 100 percent.[85]

6.2.3.2 Revisions to General Reliability Procedures

Multiple processes were used to monitor the Two Bags Task coding operation, including double-coding of 5 percent of each coder's cases and examining coding timing data, the distribution of scores, and reliability data. Weekly reports of coding timing and reliability data distributions by scale indicated that the average time required to code a tape varied more than would be expected from week to week, although it was decreasing over time. Additionally, these data suggested there was variability in the reliability among the coding staff and across the

[84] The exact number of cases was determined by the standard coder, generally in consultation with the coding task leader, to provide an adequate number of cases to identify the source of the field coder's differing scores and to use for corrective feedback. In some instances this could be done with fewer cases while in others, more cases were required for the field coder to demonstrate to the standard coder an accurate understanding of the scale and scoring for cases.

[85] Because agreement was defined as being within 1 point, the standard coder's scores could be different from the field coder's scores even when percent agreement was 100 percent.

coding scales. Variability in these indicators, where limited variability would be expected, eventually prompted further analysis of the ongoing reliability of DVDs already coded for the task. Part of this process included an assessment of the current reliability of the DVDs coded by field coders. While the existing data were being examined, field coders stopped coding the Two Bags Task.[86] On April 10 and 11, 2006, the coding task leader led a refresher training for the nine field coders remaining on the project. During training, the coding task leader stressed the importance of using the entire range of each scale, as applicable. Review of DVDs coded prior to the refresher training had suggested that there was a tendency among coders to score only 1s and 2s on the negative scales (particularly detachment and negative regard). The coding task leader asked the coders to pay more attention to these scales and to consult their manuals frequently. Additionally, the coding task leader gave some examples of behaviors associated with each point on each scale. Lastly, the field coders were given the opportunity to ask questions. Once refresher training was completed, the field coders' reliability was assessed by comparison against standard coder scores on five cases. Despite the refresher training, most of the coders were still less than 80 percent reliable with the standard coders. Consequently, it was decided to temporarily cease coding of Two Bags Task DVDs and engage in a three-step approach for verifying coder reliability.

The three-step approach was designed with a goal of coding the preschool Two Bags videotapes so that the resulting data were reliable and consistent with the ECLS-B 2-year Two Bags data, which in turn were consistent with the EHS CUTC data. In Step 1, standard coders underwent refresher training and were recertified to demonstrate 90 percent agreement within 1 point, for each scale, against 10 cases drawn from the EHS study. Step 2 required that standard coders recode 20 percent of the previously coded DVDs for each field coder from each 2-week coding period (for cases initially coded between September 2005 and January 2006). Reliability for the standard coders in Step 2 was maintained through weekly reliability checks, which are explained in more detail in the following sections. Step 3 involved adopting a new method of ensuring ongoing reliability and applying it during the coding of the remainder of the DVDs and to those determined to be unreliable during Step 2.

Step 1: Determining Current Standard Coder Reliability. Step 1 involved an assessment of the reliability of the standard coders. On August 24, 2006, three standard coders were retrained on the Two Bags Task scoring procedures using the EHS tapes and training materials previously used by Dr. Fauth.[87] The standard coders were then required to demonstrate reliability with CUTC videotaped interactions using the following procedure: From the interactions available on the videotapes provided by Dr. Fauth, 10 cases were randomly selected for use in

[86] As noted earlier, during this time field coders were allowed to complete coding of other preschool wave items, including *Let's Tell Stories* and the fine motor drawings. As a result of the reduced work load, however, some coders left the project during this time.

[87] By this point in the study, only one of the original standard coders remained at RTI. The other two standard coders were selected from the original field coders based on their demonstrated high levels of reliability during the initial period of coding activity.

testing the standard coders against the EHS standards. Each of the standard coders watched and rated each of these interactions independently. Once all 10 cases had been completed, each standard coder's scores were compared against the scores provided by Dr. Fauth for each interaction. Agreement was defined as either an exact matching score or a score within 1 point of that provided by Dr. Fauth. All three standard coders showed 100 percent agreement (exact match or within 1 point) against Dr. Fauth's scores on all scales. As a final training step, the coders discussed the scales on which they had any difference in scoring against the provided scores.

Step 2: Determining Reliability of Coded Tapes Over Time. Once their reliability against the EHS scores was demonstrated, standard coders began work on Step 2 of the process. Step 2 required that standard coders recode 20 percent of the previously coded DVDs for each field coder from each 2-week coding period for cases initially coded between September 2005 and January 2006. Within a given 2-week coding period, for a given coder, standard coders recoded 20 percent of tapes, with a minimum of 10 tapes. When there were fewer than 10 tapes within a given 2-week coding period for a given coder, the standard coder recoded all tapes for that coding period for the given coder. (For example, if there were 80 tapes in a 2-week period for a given coder, the standard coder recoded 16 tapes. If there were 40 tapes in a 2-week period for a given coder, the standard coder recoded 10 tapes. If there were 8 tapes within a 2-week period for a given coder, the standard coder recoded all of those 8 tapes.)

For the recoded tapes, percentage agreement between the original coding, done prior to and during Step 1, and the standard coders' recoding (within 1 point) was calculated for each scale. If the coder's original scores were within 1 point of the standard coder's scores at least 85 percent of the time per scale, for all scales, the coder's work for that 2-week period was considered reliable. If, however, the coder's percent agreement was below 85 percent on any scale, the coder's work during that time period was considered to be unreliable, and all cases (except those scored by the standard coder as part of Step 2) were recoded by a new set of certified field coders as part of Step 3 (described below). If a coder's work was determined to be unreliable for three consecutive coding periods, all of the coder's cases in those periods and all subsequent periods (except for cases already coded by standard coders) moved to Step 3 and were recoded by a new set of certified field coders. This process resulted in scores for every case scored being completed by a field coder whose reliability had been monitored or verified against scores from a standard coder or by a standard coder whose reliability had been certified against established EHS scores.

Because Step 2 was conducted over several months, it was important to ensure that all three standard coders remained consistent in application of the EHS standard to ECLS-B cases over the course of the recoding period. This goal was achieved through the weekly scoring of a common set of cases that all three standard coders independently scored. The standard coders then discussed the scores assigned for each case and adjudicated any differences between scores assigned (note that in this process, scores would be considered different if they did not match

exactly). The result of this process was an increasing number of cases with standard scores for each scale that was used to assess field coder reliability. In addition, the process of scoring and adjudication served to maintain the standard coders' consistent application of scores throughout the duration of their work. Moreover, because this was reflected in the scores used to monitor field coders' reliability, this process also helped to maintain field coders' consistent application of the scores over time. Finally, if any standard coder was inactive for 5 or more business days, that coder could not resume production until he or she had completed and demonstrated agreement (within 1 point) with one of the EHS cases used in initial certification.

Table 51 summarizes the results of the reliability assessment for cases initially completed by field coders between September 2005 and January 2006, as examined under Step 2, for each subscale and the overall mean reliability per scale. Note that scores deemed reliable under Step 2 were retained on the data file; the work of coders who were shown not to be reliable under Step 2 was subsequently recoded under Step 3, and are not included in table 51.

Table 51. Average reliability (percent agreement) for subscales of the Two Bags Task for the ECLS-B preschool data collection, initial coding: 2005–06

Scale	Average percent agreement	Range of percent agreement
Parental emotional supportiveness	97.1	85–100
Parental stimulation of cognitive development	96.7	85–100
Parental intrusiveness	94.1	85–100
Parental negative regard	96.4	85–100
Parental detachment	96.0	85–100
Child engagement	95.4	85–100
Child quality of play	93.3	85–100
Child negativity	97.0	85–100

NOTE: Data are from the 13 original field coders who contributed scores to the final data set. This table only includes data for work determined to be reliable, and reflects the data included on the data file. Work that was determined to be nonreliable was subsequently recoded.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Step 3: Training Field Coders Coding New Tapes and Those Determined to Be Unreliable During Step 2 Using EHS Standards. After procedures to assess coder drift (i.e., Step 2) had been underway for 10 weeks, Step 3 was initiated. Due to attrition among field coders, a second set of 10 field coders was trained at RTI on November 18 and 19, 2006, by the coding task leader who had conducted the previous training in September 2005 (one of these coders left the project shortly after training and prior to certification, leaving nine field coders.). The coding task leader was no longer employed at RTI but nevertheless agreed to come to RTI for the distinct purpose of leading this training.[88] As the initial task manager and one of the initial standard coders, this individual had been trained by Dr. Fauth at CUTC at the outset of the study and employed the same agenda and approach used during the training for the initial field coders. The coding task leader provided an overview of the coding task, reviewed the rating scales, and

[88] When the initial coding task leader left the project, a new RTI task leader was selected to manage this effort, with scientific oversight provided by the study PI.

provided individual practice and group discussions. The training was informed by the standard coders' experience in scoring cases.

The training of field coders for Step 3 differed from the first field coder training in several respects. First, because the trainer was unavailable to lead certification, the certification exercises were conducted and monitored by the current standard coders, under the oversight of the current coding task leader and principal investigator (PI). Certification was conducted using the original set of EHS cases (sets of five were used to allow for multiple certification efforts, if necessary). This modification was intended to reinforce the alignment between the work done on the EHS study and the ECLS-B. Second, the criterion for certification on the Two Bags Task was modified so that coders were required to achieve 80 percent agreement on each scale (based on five certification cases), rather than an average of 80 percent. This modification was made to further enhance the initial reliability of this set of coders and to protect against possible coder drift on any one scale (i.e., a coder "drifting" from agreement with the standard over time).

As part of the certification process, immediately following training, all of the coders had to score a set of Two Bags Task cases previously scored by the current standard coders. As mentioned above, to be certified, the coders had to demonstrate an 80 percent agreement rate with the assigned standard coder scores on each scale (i.e., each coder had to demonstrate agreement [within 1 point] with standard coder scores on four out of the five ECLS cases for each scale). Note that these criteria required 80 percent agreement on each scale, rather than an average of 80 percent agreement across all scales. This was achieved over three certification rounds,[89] with five coders passing certification on the first round. In this way, initial reliability for the new coders was established. Table 52 provides the initial reliability for the nine field coders trained as part of Step 3.

[89] Each certification round required the coder to code 10 ECLS-B Two Bags tapes and match the standard scores assigned to each case (within 1 point) for 8 out of 10 cases for each scale (i.e., an 80 percent agreement rate).

Table 52. Percent agreement for the Two Bags Task Step 3 field coders, by scale, ECLS-B preschool data collection initial certification: 2005–06

Scale	*n*	Average percent agreement	Range of percent agreement for passing scores
Parental emotional supportiveness	9	100.0	100
Parental stimulation of cognitive development	9	93.3	80–100
Parental intrusiveness	9	86.7	80–100
Parental negative regard	9	100.0	100
Parental detachment	9	97.8	80–100
Child engagement	9	97.8	80–100
Child quality of play	9	97.8	80–100
Child negativity	9	95.6	80–100

NOTE: Certification scores from the round in which the coder passed certification. Data are from the nine Step 3 coders who contributed scores to the final data set.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Ongoing reliability for Step 3 coders was assessed through their scoring of a reliability case at the end of each week (see table 53 for reliability on these cases). These cases had been scored through consensus by standard coders as part of Step 2. Coders who were within 1 point of the consensus score per scale on all scales were considered reliable and continued scoring cases. Coders who did not meet this criterion received direct feedback and additional instruction (if necessary) and scored a second reliability case. If they demonstrated agreement (within 1 point) on all scales on this second case, they continued to score cases as before. If they did not meet the criterion on this second reliability case, all of the coder's work completed between the previous reliability check and the one he or she failed (i.e., the week prior to failing the reliability check) was assumed to be unreliable and the coder ceased scoring live cases. Cases completed during this period of assumed unreliability were returned to the coding queue for recoding by a coder deemed to be reliable. Next, the coder in question reviewed additional descriptions of the coding scales and viewed videotapes from the November 2006 training, with particular emphasis on those scales for which the coder was unable to match the reliability case scores. The coder's reliability was then assessed at the next regularly occurring reliability check. If the coder met criteria, she resumed production coding, if not, the steps noted above were again taken. This is consistent with the approach taken during the 2-year wave of data collection when coders were found to be unreliable during a period of coding activity.

Table 53. Average reliability (percent agreement) against weekly reliability cases for subscales of the Two Bags Task for the ECLS-B preschool data collection, Step 3 coders: 2005–06

Scale	Average percent agreement	Range of percent agreement
Parental emotional supportiveness	96.9	92–100
Parental stimulation of cognitive development	96.4	89–100
Parental intrusiveness	96.2	85–100
Parental negative regard	99.0	91–100
Parental detachment	99.2	93–100
Child engagement	95.4	89–100
Child quality of play	99.7	88–100
Child negativity	99.1	92–100

NOTE: Data are from the nine Step 3 coders who contributed scores to the final data set.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Thus, reliability was closely monitored during Step 3. Step 3 coders completed scoring preschool Two Bags DVDs that had not yet been scored and also cases that had been scored by the original field coders but were subsequently identified during Step 2 as needing rescoring (see above).

Sources of Final Scores. Coders were able to complete scoring on about 7,650 preschool cases (93.8 percent of cases received).[90] Of these, approximately 1,450 cases (19.2 percent) were scored by standard coders; approximately 2,450 (32.1 percent) had been completed by the initial set of field coders and were verified as reliable through Step 2; and approximately 3,700 (48.7 percent) were completed by Step 3 field coders. The Step 3 coders completed approximately 2,300 cases (30.3 percent of all coded cases) that were rescored from initial field coder work that had been found to be unreliable in addition to approximately 600 cases (7.8 percent of the total coded) that were rescored without a determination concerning initial reliability.[91] Finally, approximately 800 cases (10.8 percent of the total) received following the end of the initial coding period were scored in Step 3. In all cases, regardless of percent agreement score, the standard coder's scores, when available, were assigned to double-coded cases, thereby replacing those of the field coders. The sources of scores for the Two Bags Task are shown in table 54.

[90] Approximately 6 percent of the received tapes could not be coded due to various reasons including poor recordings, the use of languages other than English and Spanish, or the duration of the task being too brief to code (i.e., less than 5 minutes).

[91] To facilitate completion of coding within study deadlines, these cases were recoded by Step 3 coders without verification by standard coders in Step 2. It should be noted that these cases were of unknown reliability, but coders completing Step 3 were functioning with known reliability, so their scores were preferable to those from initial coders without verification.

Table 54. Sources of Two Bags Task scores, preschool data collection: 2005–06

Source of score	Number	Percent of all cases	Percent of all scored cases
Total cases	8,150	†	†
Not codeable due to recording errors	400	4.9	†
Total number of cases with readable DVDs	7,750	95.1	†
Not scored due to language other than English or Spanish	100	1.3	†
Total cases scored	7,650	93.8	†
Cases scored by the standard coders	1,450	†	19.2
Cases scored by field coders and determined to be reliable in Step 2	2,450	†	32.1
Cases rescored by field coders in Step 3 due to low initial coder reliability	2,300	†	30.3
Cases rescored by field coders in Step 3 without passing through Step 2	600	†	7.6
Cases scored in Step 3, not previously scored	800	†	10.8

† Not applicable.

NOTE: The "Number" of cases has been rounded to the nearest 50. Details may not sum to the total due to rounding. "Total cases" refers to the number of cases for which the task was administered and consent was given for recording the interaction. "Total number of cases with readable DVDs" indicates the number of cases for which a DVD was received and able to be read by coders' computers. A case was identified as "not codeable due to recording errors" if a technical problem precluded the scoring of the case. A case was identified as "not scored due to language other than English or Spanish" if the use of a language other than English or Spanish precluded a coder from scoring the case. Each case could be identified with either, or both, of these problems, precluding it from being scored.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

6.2.4 Two Bags Task Data in the Preschool Collection

Table 55 provides descriptive statistics for the Two Bags Task rating scales.

Table 55. Weighted means and standard deviations for the Two Bags Task rating scales in the ECLS-B preschool data collection: 2005–06

Scale	*n*	Mean	*SD*	*Range*
Parental emotional supportiveness	7,600	4.42	0.92	1–7
Parental stimulation of cognitive development	7,600	4.19	0.97	1–7
Parental intrusiveness	7,600	1.53	0.87	1–6
Parental negative regard	7,600	1.19	0.49	1–6
Parental detachment	7,600	1.31	0.69	1–7
Child engagement	7,600	4.48	0.89	1–7
Child quality of play	7,600	4.04	0.89	1–7
Child negativity	7,600	1.35	0.74	1–7

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. A total of approximately 7,650 cases had scores on at least 1 scale. In some cases, one or more scales could not be coded due to problems with the recording during sections of the interaction. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data. Only those cases with a valid weight are included in the table. *SD* = standard deviation.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Although individual scale scores are provided to indicate independent constructs, in practice the scoring guidelines provided some potential for overlap during the scoring. For example, it is reasonable to expect that parental intrusiveness and parental supportiveness would be negatively related, if at all. However, the relative independence of the scales can be more directly addressed by examining zero-order correlations among the individual rating scale scores. These correlations are shown in table 56.

Table 56. Pearson correlation coefficients across scales, ECLS-B preschool data collection: 2005–06

Scale	Parental emotional supportiveness	Parental stimulation	Parental intrusive-ness	Parental negative regard	Parental detach-ment	Child engage-ment	Child quality of play
Parental emotional supportiveness							
Parental stimulation of cognitive development	.61						
Parental intrusiveness	-.15	-.09					
Parental negative regard	-.25	-.15	.38				
Parental detachment	-.35	-.31	.19	.27			
Child engagement	.46	.42	-.20	-.15	-.18		
Child quality of play	.40	.51	-.05	-.09	-.12	.48	
Child negativity	-.14	-.12	.57	.34	.14	-.31	-.11

NOTE: Correlations based upon a sample of approximately 7,600 cases. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data. Only those cases with a valid weight are included in the table. All correlations $p < .001$.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

In addition to scores on the interaction scales, several additional variables related to the Two Bags Task were coded that may be of interest to analysts (see table 57). These variables include the language the parent used with the child during the activity, whether the parent and child had ever read the study book before, and if so, how often.[92] For twin cases, there is a contextual variable (C31ST2ND) for twin pairs that indicates for each twin whether he or she was the first or second to complete the Two Bags Task.

Also, for each interaction, the total duration was recorded by coders, starting from the time the field interviewer completed instructions to the parent for implementing the task and handed the first bag to the parent and ending when the field interviewer indicated that the task was complete (after 10 minutes of interaction) or when the parent and child had placed all items back into the bags (if this occurred prior to 10 minutes' duration). The protocol allowed for 10 minutes of interaction to be recorded; however, in practice, the interaction often exceeded the 10-minute target. When the recordings were coded, however, only the first 10 minutes of interaction were scored. The duration of the interaction is captured with two variables in the data file: C3LGINTM gives the number of minutes and C3LGINTS gives the number of seconds. To obtain the exact duration of the interaction, these two variables must be combined (e.g.,

[92] During the preschool wave, interactions were coded only if they were primarily recorded in English or Spanish. Cases that were substantially non-English and non-Spanish were not coded.

C3LGINTM = 7 and C3LGINTS = 30 corresponds to a duration of 7:30). Coders also recorded the time that the dyad spent with the book. This duration was measured from the time the parent removed the book from the first bag, and ended when the book was put back in the bag, or set aside as the dyad moved to the second bag. In cases where the parent put the book down but later returned to it, both segments involving the book were combined when calculating the duration of time spent with the book. As with the total duration of the interaction, time spent with the book is captured in two variables: C3LGBRDM provides the number of minutes and C3LGBRDS provides the number of seconds. To determine the exact duration of book reading, these two variables must be combined.

Table 57. Variable name, label, description, and scoring for Two Bags Task variables, preschool data collection: 2005–06

Variable name	Item description	Values
C3EMOSPT	Parent Behavior—Emotional Supportiveness	1–7
C3COGDEV	Parent Behavior—Stimulate Cognitive Development	1–7
C3NTRUSV	Parent Behavior—Intrusiveness	1–7
C3NEGRGD	Parent Behavior—Negative Regard	1–7
C3DETACH	Parent Behavior—Detachment	1–7
C3ENGPRT	Child Behavior—Engagement with Parent	1–7
C3QUALTY	Child Behavior—Quality of Play	1–7
C3NEGPRT	Child Behavior—Negative with Parent	1–7
X3TBLNG	Language of Two Bags Task	1= English; 2 = Spanish; 3 = other
X3OFTNRD	How many times read book before	0 = have never read the book before; 1 = once or occasionally; 2 = a couple of times or two times; 3 = a few times or 3–6 times; 4 = several times or 7–10 times; 5 = many times or more than 10 times
C3READBK	Had book been read before	1 = yes; 2 = no
C31ST2ND	If twins, was this twin the first or the second	1 = first; 2 = second; 3 = this was not a twin case
C3LGINTM	Length of interaction, minutes	5–27
C3LGINTS	Length of interaction, seconds	0–59
C3LGBRDM	Length of book reading, minutes	0–14
C3LGBRDS	Length of book reading, seconds	0–59
C3PLACE	Where did the task occur	1 = table; 2 = floor; 3 = some other place
C3DSTRCT	Distractions by others present	1 = yes; 2 = no

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Finally, because the Two Bags Task involves both book reading and play with objects, the setting in which the task occurred may be of interest to users. The field interviewers were instructed to complete this task with parent and child seated at a table, but the protocol allowed the task to be completed in any location in which the parent and child could be seated together. As a result, some of the interactions took place with the parent and child at a table and others took place with the parent and child sitting together with a table (e.g., coffee table, folding table) nearby. A variable on the data file (C3PLACE) indicates whether the task was completed while seated at a table, while sitting on the floor, or at some other location in the home. An additional

variable (C3DSTRCT) is provided to indicate whether, during the interaction, others were present who were potentially distracting to the parent or child.

6.2.5 Comparison of the Preschool Two Bags Task with the 2-Year Two Bags Task

The Two Bags Task administered during the preschool wave was largely similar to the task as it was administered during the 2-year data collection. The procedure at the preschool wave was modified slightly to be developmentally appropriate for the children when they were older. During both waves, bag 1 contained a children's book, but it was different in the two waves. At 2 years, the book was *Good Night, Gorilla* by Peggy Rathmann (1994), and at preschool it was *Corduroy* by Don Freeman (1968). The toys provided in the second bag also were changed between waves, with the preschool wave replacing the set of dishes used at 2 years with two containers of Play-Doh®, a rolling pin, and two cookie cutters. At 2 years, the interaction was videotaped on 8 mm (Hi-8 format) tapes using a tripod-mounted camcorder; in the preschool wave, the field interviewer recorded the interaction on a DVD, again using a tripod-mounted camcorder in the home.

The coding scales were modified to remain developmentally appropriate and also in response to findings from the 2-year administration. The 2-year parent rating scales of Sensitivity and Positive Regard were found to be highly correlated at 2 years. Consequently, they were merged into one scale for the preschool wave, named Emotional Supportiveness.[93] Also for the preschool wave, the child scale Sustain Attention was not scored, because it was no longer developmentally appropriate. It was replaced with the child scale Quality of Play. Otherwise, the scales scored were common to both waves, with some modification to the description and scoring guidelines to make them more appropriate for the older sample.

6.2.6 Correlations of Preschool and 2-Year Two Bags Task Scores

The general commonality of specific coding scales between the 2-year and preschool waves allows for an examination of the correlations between common scales across the two data collection waves. These correlations (shown in table 58) must be considered with some caution, however, given the long interval between observations (approximately 2 years), the possibility that the adult may be different in the two interactions, and the limited variability found for some scales (e.g., parental intrusiveness, child negativity).

[93] Note that this should not be confused with the composite score developed during the 2-year data collection called Parental Supportiveness, which averages not only the 2-year Sensitivity and Positive Regard scales, but also includes the Cognitive Stimulation scale.

Table 58. Weighted correlations between common preschool and 2-year Two Bags Task rating scale scores: 2003–04 and 2005–06

2-year score	Preschool score	*n*	Correlation
Parental sensitivity	Parental emotional supportiveness	6,200	.31
Parental positive regard	Parental emotional supportiveness	6,200	.30
Parental stimulation of cognitive development	Parental stimulation of cognitive development	6,200	.30
Parental intrusiveness	Parental intrusiveness	6,200	.12
Parental negative regard	Parental negative regard	6,200	.16
Parental detachment	Parental detachment	6,200	.12
Child engagement	Child engagement	6,200	.20
Child sustained attention	Child quality of play	6,200	.23
Child negativity	Child negativity	6,200	.09

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data. The weight corresponds to the latest wave represented by the correlation. Only those cases with a valid weight are included in the table.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 2-year (2003–04) and preschool (2005–06) data collections.

6.3 Reading Aloud Profile–Together Coding in the Preschool Data Collection

In addition to the Two Bags Task coding discussed above, a subsample (n = 800) of the preschool Two Bags Task DVDs also was coded using the Reading Aloud Profile–Together (RAPT) coding scheme to provide detailed information about parents' and children's behaviors while engaged in a joint book reading activity. The RAPT data can be used to examine whether joint book reading behaviors of parents and children vary by family and child characteristics and whether joint book reading behaviors relate to children's early reading competency at preschool and upon entry to kindergarten, among other things.

The RAPT was developed for the Even Start Classroom Literacy Interventions and Outcomes Study (CLIO) in 2004 as an adjunct to the Read Aloud Profile (RAP), which is part of the Observation Measure of Language and Literacy Instruction (OMLIT) observation system (Goodson et al. 2004). RAP had focused solely on the behavior of the adult when he or she is reading aloud with children; RAPT added codes for discrete child behaviors during the parent-child interaction task. The development of RAPT was an advance in the measurement of book reading behaviors; most joint book reading coding schemes focus on parent behavior and assess child engagement only globally. Like its predecessor, RAP, RAPT was designed to capture behaviors related to the major domains of early literacy: comprehension/higher-order thinking (e.g., new vocabulary, information about the content of the text, linking the meaning of the text to the child's own experience, and review of the text/linked activities to build on understanding of the text), use of open-ended questions, print motivation, print knowledge, and phonological awareness. The child codes on the RAPT are designed to capture (a) evidence of the child's interest and motivation in reading, (b) the child's engagement with print, (c) the child's attempts to build comprehension through questioning, and (d) the child's use of oral language to communicate thinking. The behavior codes for both parent and child are coded dichotomously to

indicate presence or absence of specific behaviors thought to be related to the major domains of early literacy.

To code these targeted behaviors, the RAPT divides the joint book reading activity into three distinct phases: (1) activity before reading the book, (2) activity during book reading, and (3) activity after reading the book. Activity before reading the book includes all book-related discussion and activity prior to beginning the text of the story (discussion about the book itself, the cover, the title pages, etc.). The second phase of the activity, during book reading, begins once the story text has begun. When the story text is completed and the dyad is no longer discussing the last page, any further discussion about the book or the story is coded as part of activity after reading.

In addition to coding discrete behaviors exhibited during the three phases of the book reading task, the RAPT coding scheme also contains three global quality indicators that take into account interaction across all three phases: story-related vocabulary, use of open-ended questions, and depth of parent-child discussion. These global quality indicators are five-point Likert scales with higher scores indicating more extensive demonstration of the indicator.

To limit costs, a simple random subsample of approximately 800 of the Two Bags Task DVDs collected during the ECLS-B preschool wave was coded using the RAPT coding scheme. Because the RAPT subsample was based on a simple random selection from the ECLS-B sample, which included multiple oversamples, this subsample design, in effect, results in a subsample with the same pattern of oversampling present in the main sample. The size of the subsample allowed precision in estimates to detect differences at alpha 0.05 with 80 percent power.

Because the ECLS-B data are gathered from a sample, rather than the population, the data must be weighted appropriately to obtain nationally representative estimates. Further, because the sample is a clustered, multistage sample, standard errors must be adjusted appropriately to conduct significance tests. RAPT data can be weighted using a main sample weight (i.e., W3R0). To obtain accurate population estimates the sample weight must be adjusted by multiplying it by the inverse of the probability of selection (for example, W3R0 * [8,900/800]). For most analyses, users do not need to adjust the replicate weights in the same manner as the main sample weight. The one exception is when producing standard errors for population sizes (i.e., *n*). To produce appropriate standard errors for population sizes, users should multiply the replicate weights by the inverse of the probability of selection, as is done for the main sample weight. For more information on main sample weight selection and standard error adjustment, see chapter 5 of the ECLS-B Kindergarten 2006 and 2007 Data File User's Manual (Snow et al. 2009).

6.3.1 RAPT Coding

As noted above, the RAPT divides the joint book reading activity into three distinct phases: (1) activity before reading the book, or prereading; (2) activity during book reading, or

during reading; and (3) activity after reading the book, or postreading. It also provides several global measures of book reading quality. Variables derived from the RAPT coding are described in sections 6.3.2.1 through 6.3.2.3. Frequency distributions for all RAPT variables are provided in table 59. Descriptive statistics for the global quality indicators are shown in table 60.

6.3.1.1 Coding Prereading Activity

In the ECLS-B administration, prereading began when the parent or child removed the book from the first bag. Both the parent's and the child's behavior were noted. The parent was scored (yes/no) for engaging in any of 10 behaviors: capturing child's attention, ensuring child comfort, verbally directing child to features of the book, pointing to features of the book, noting different letters or letter sounds, reminding the child of other similar books, responding to the child's questions, asking closed-ended questions, asking open-ended questions, and relating the book to the child's experience. The parent needed to demonstrate the behavior only once to be scored as demonstrating the behavior; frequency of behaviors was not captured. A code is available if the parent did not evidence any of these behaviors. The child was scored (yes/no) for demonstrating any of six behaviors: showing interest in the book, verbally answering the parent's questions about the book, pointing out features of the book, asking questions about the book, expanding on the parent's comments regarding the book, and talking about the book's story line. There is a code to indicate if the child did not evidence any of the above behaviors; frequency of behaviors was not captured.

6.3.1.2 Coding Reading Activity

Once the dyad began reading the story text or talking about the illustrations that accompanied the story text, the dyad moved into the during-reading phase of the activity. Both the parent and the child were scored (yes/no) for engaging in as many as 12 different behaviors. Only the presence or absence of the behavior was noted; frequency was not captured. However, if the dyad chose to read the book a second time, the same 12 behaviors were targeted and, if evidenced, were scored and included in the data file (with a corresponding code noting the occurrence of a second read).

During reading, the parent was scored (yes/no) for engaging in any of the following behaviors: tracking print, using gestures or dramatic voices, directing the child's attention to the illustrations, asking the child story-related closed-ended questions, expanding on the story or on the child's comments, commenting on letters and sounds, highlighting new vocabulary, asking recall questions, relating the story to the child's experience, asking story-related open-ended questions, and having the child join in the reading. A code is available to indicate that the parent did not evidence any of the above-mentioned behaviors during the book reading activity. The child was scored (yes/no) for engaging in several possible behaviors: attending to the story, losing interest, verbally responding to questions about the book, pointing to things in the book, labeling (i.e., naming objects in the illustrations), acting out the story, repeating words, relating the story to his or her own life, making comments about the story or something the parent said,

asking questions about the story or something the parent said, and trying to read the book. A code is available to indicate that the child did not evidence any of the above-listed behaviors.

6.3.1.3 Coding Postreading Activity

After completing the story, any further book discussion by the dyad was coded as postreading activity. Because the Two Bags Task included a second bag with which the dyad could interact, postreading activity was not done extensively, although many parents were scored (yes/no) for asking the child if he or she liked the book. However, parents also could be scored (yes/no) in postreading for allowing the child to look at the book, answering questions from the child regarding the story or story topic, expanding on the child's comments about the book, reviewing vocabulary in the book, asking the child to recall parts of the book, asking story-related open-ended questions, relating the story to the child's experience, and summarizing the story either with or without child participation. The child was scored (yes/no) during post-reading for asking to read the book again, responding to parent questions, commenting on the book, asking questions about the story, and trying to read the book. A code is available to note if no postreading activities took place.

6.3.1.4 Other Data

Ancillary information was also coded, including the length of interaction, the language spoken by the dyad (English, Spanish, English and Spanish, or English plus another language), the language of the book (English or Spanish), and whether the dyad read through the book a second time. Coders recorded reasons a videotape was uncodeable and noted any special circumstances that may have interfered with coding (see section 6.3.4).

Table 59. Frequency distribution for RAPT parent and child behaviors

Variable name	Variable description	Frequency percent for "yes" value
Z3BPCOMF	Before Reading: Parent makes child comfortable	4.6
Z3BPATTN	Before Reading: Parent captures child's attention	75.7
Z3BPLABL	Before Reading: Parent labels parts of book	76.9
Z3BPPNTS	Before Reading: Parent points to book parts	36.3
Z3BPPHON	Before Reading: Parent identifies sounds/letters	0.9
Z3BPREMD	Before Reading: Parent reminds child of similar books	14.1
Z3BPRSPD	Before Reading: Parent responds to child's questions	5.2
Z3BPCLQS	Before Reading: Parent asks closed-ended questions	43.2
Z3BPRLAT	Before Reading: Parent relates book to child's experiences	5.5
Z3BPOPQS	Before Reading: Parent asks open-ended questions	3.3
Z3BPNONE	Before Reading: No parent prereading activity	4.2
Z3BCINTR	Before Reading: Child shows interest in book	57.0
Z3BCRSPD	Before Reading: Child responds to parent's questions	38.2
Z3BCPNTS	Before Reading: Child tells parent about book	16.7
Z3BCASKQ	Before Reading: Child asks questions about book	4.9
Z3BCEXPD	Before Reading: Child expands on parent's comments	1.9
Z3BCTLST	Before Reading: Child tells parent about story	2.8
Z3BCNONE	Before Reading: No child prereading activity	31.0
Z3DPTRK	During Reading: Parent: tracks print	36.8
Z3DPACT	During Reading: Parent: acts out story	43.1
Z3DPPIC	During Reading: Parent: directs child to pictures	84.8
Z3DPCLQ	During Reading: Parent: asks closed-ended questions	76.9
Z3DPSTY	During Reading: Parent: expands on story	63.5
Z3DPANS	During Reading: Parent: answers child's questions	42.2
Z3DPLTR	During Reading: Parent: highlights letters	1.0
Z3DPVOC	During Reading: Parent: highlights new vocabulary	13.9
Z3DPREC	During Reading: Parent: asks child to remember back	10.3
Z3DPRLT	During Reading: Parent: relates book to child's experiences	35.1
Z3DPOPQ	During Reading: Parent: asks open-ended questions	29.4
Z3DPCRD	During Reading: Parent: has child read text	10.6
Z3DPNON	During Reading: No parent during-reading activity	5.3
Z3DCATT	During Reading: Child attends to story	91.9
Z3DCRSP	During Reading: Child responds to parent's questions	61.9
Z3DCPNT	During Reading: Child points to pictures/text	51.9
Z3DCLBL	During Reading: Child labels pictures	36.1
Z3DCRPT	During Reading: Child repeats words/story	24.2
Z3DCACT	During Reading: Child acts out parts of book	10.4
Z3DCREL	During Reading: Child relates story to own experiences	11.5
Z3DCCOM	During Reading: Child comments on activity	53.9
Z3DCQST	During Reading: Child asks questions	32.7
Z3DCTRN	During Reading: Child explores book on own	3.3
Z3DCTLL	During Reading: Child read book/tells story	10.6
Z3DCLOS	During Reading: Child loses interest	49.6
Z3DCNON	During Reading: No child during-reading activity	1.4
Z3APINT	After Reading: Parent asks if child liked book	34.4
Z3APLIK	After Reading: Parent lets child to look at book	6.9

See notes at end of table.

Table 59. Frequency distribution for RAPT parent and child behaviors—Continued

Variable name	Variable description	Frequency percent for "yes" value
Z3APANS	After Reading: Parent answers questions about book	2.8
Z3APRES	After Reading: Parent responds to child's comments	4.4
Z3APVOC	After Reading: Parent reviews vocabulary	0.0
Z3APREC	After Reading: Parent asks for recall of book	6.7
Z3APREL	After Reading: Parent asks questions related to child's experiences	5.1
Z3APOPQS	After Reading: Parent asks open-ended questions	2.2
Z3APSMWO	After Reading: Parent summarizes book without child's involvement	1.5
Z3APSMWH	After Reading: Parent summarizes book with child's involvement	1.9
Z3APNONE	After Reading: No parent post-reading activity	54.2
Z3ACRERD	After Reading: Child asks to read book again	3.3
Z3ACRESP	After Reading: Child responds to questions about book	10.8
Z3ACCOMT	After Reading: Child comments on story/illustrations	7.0
Z3ACASKQ	After Reading: Child asks questions about book	3.5
Z3ACREAD	After Reading: Child tries to read book on own	0.5
Z3ACNONE	After Reading: No child postreading activity	79.3

NOTE: About 100 of the 800 cases sampled were not codeable for various reasons. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data. Only cases with a valid weight are included in the estimates. Percents are based on cases with valid responses.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Table 60. RAPT quality indicator summary statistics

Variable name	Variable description	*n*	Mean	*SD*	Range
Z3VOCAB	Parent's attention to vocabulary	700	1.12	.353	1–3
Z3OPENQS	Parent's use of open-ended questions	700	1.43	.841	1–5
Z3DISCUS	Depth of parent/child discussion	700	1.21	.682	1–5

NOTE: The sample size (*n*) has been rounded to the nearest 50. About 100 of the approximately 800 cases sampled were not codeable for various reasons. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data. Only those cases with a valid weight are included in the table.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

6.3.2 RAPT Coder Training

RAPT training was held October 3–6, 2006, at the Rockville Institute in Rockville, Maryland. Lisa Pletcher, an expert in video coding and the head of the CLIO coding laboratory, led training. The ECLS-B RAPT project leader and the ECLS-B RAPT lead coder attended the training, along with four other ECLS-B staff assigned to the RAPT task. One staffer was fluent in Spanish and was expected to code the Spanish ECLS-B RAPT DVDs.

Training began with a discussion of observational coding and an observational exercise. Each attendee was given a training manual, a copy of the books used for the CLIO study, and coding sheets. The coding manual was reviewed, and the remainder of training comprised viewing and coding primarily CLIO video recordings, along with a handful of ECLS-B Two Bags DVDs. During training, video recordings were viewed as a group. However, each trainee individually coded each video recording without consulting or discussing the video recording

with any of the other trainees. Each video recording contained one parent-child book reading interaction. At the conclusion of each video, the lead trainer discussed assigned codes with ECLS-B coders. In this way, coders learned to apply the RAPT coding scheme consistent with the CLIO coding team.

Certification was held on the last day of training. Coders viewed 10 recordings and coded them independently. Six of these video recordings were from the CLIO study and four were from the ECLS-B. After each video recording, the lead coder reviewed the standard codes and each trainee noted any discrepancies on his or her coding sheet. A coder was certified as reliable if the rate of agreement between the coder and standard was within two codes for each member of the dyad within each of three phases of reading activity.[94] The quality indicators, however, required coders to score within one point of the standard to be deemed reliable. Coders were certified if they demonstrated 90 percent reliability or higher on each member of the dyad in each phase of reading activity, and on the quality indicators. This meant that for each member of the dyad, in each phase of the reading activity, a coder had to come within 2 points of the standard for 9 out of the 10 certification cases and had to come within 1 point on each of the three quality indicators for 9 out of the 10 certification cases. All coders were able to demonstrate 90 percent reliability or higher on the last day of training.[95]

Three ECLS-B RAPT coders left the project at the end of January 2007. The ECLS-B lead coder trained three new coders at the National Center for Education Statistics in February 2007. This training followed a similar format to the initial training.[96] Each trainee individually rated DVD recordings. Each DVD held one parent-child interaction. At the conclusion of each interaction, the trainer discussed the correct codes with the coders. When coders felt ready, they recoded 10 previously coded video recordings for certification. All coders were required to demonstrate 90 percent reliability with the original codes, using the same standards discussed above. All coders were certified by the end of February 2007.

6.3.3 RAPT Ongoing Reliability

Observational coding of this nature is subject to coder drift (i.e., a coder "drifting" from agreement with the standard over time). To guard against this, coders were required to demonstrate reliability each week. The lead coder and the ECLS-B RAPT task leader coded several "reliability DVDs" to establish standard codes. Coders were evaluated on weekly basis[97]

[94] The RAPT coding sheet has six columns. Each column denotes a participant in the dyad (caregiver or child) for each phase of the activity (before reading, during reading, and after reading). A coder was said to be reliable if he or she were within two codes of the standard in each column.
[95] One coder was unable to attend the last day, but subsequently passed certification. To certify as reliable, this coder independently coded four ECLS-B DVDs used for certification during her absence, and six video recordings the Lead Coder had previously coded.
[96] The video recordings used for the second coder training were "reliability DVDs." Upon returning to the ECLS-B RAPT lab, the ECLS-B task leader and the ECLS-B lead coder began producing reliability DVDs. For details on the procedures for coding the reliability DVDs, please see section 6.3.3 of this report.
[97] This was typically at the beginning of their last shift each week.

on the percent agreement observed between the codes they assigned to one "reliability DVD" and the standard codes assigned to this case by the ECLS-B task leader and lead coder. If the coder was unable to demonstrate reliability (defined as being within two codes of the standard for each member of the dyad in each section of the reading activity and within one code of the standard for all three quality indicators), the coder had the opportunity to code a second reliability DVD.[98] If the coder once again failed to demonstrate 100 percent reliability, the coder immediately ceased coding. In this circumstance, DVDs coded during that week by the affected coder were checked by the lead coder for reliability and recoded if necessary. The affected coder received retraining, and required recertification before resuming coding live cases.[99] If a coder missed work during the coding period, he or she was required to demonstrate reliability before resuming live coding.

These data were analyzed for accuracy (i.e., whether the codes entered by the two coders from the hard-copy score sheets were the same) on a biweekly basis; overall, coders' data entries differed by less than 2 percent.

Fatigue can also compromise coding quality. To guard against this, no more than six DVDs were coded per day, per coder. Coders were required to take a 15-minute break every 2 hours (roughly halfway through their coding shift), with additional breaks as necessary.

Weekly team meetings also supported ongoing reliability. During these meetings, unusual circumstances were discussed to reach consensus on coding assignments. Video interactions were sometimes viewed and discussed to reach a better understanding as to how to apply the RAPT codes to the ECLS-B DVDs. Lisa Pletcher of the CLIO lab provided ongoing guidance when coding questions required greater expertise.

6.3.4 Special Codes

6.3.4.1 Uncodeable or Challenging-to-Code DVDs

Approximately 13 percent (about 100) of the RAPT DVDs sampled were uncodeable for various reasons. Most often, a DVD was uncodeable because the dyad did not spend at least 2 minutes, the minimum amount of time required to code the activity, with the book. Additionally, some DVDs were uncodeable for the RAPT because there was insufficient light during taping and the dyad was difficult to see on the videotape, because the book was out of frame, or because the dyad spoke a language other than English or Spanish. Cases that were uncodeable are included in the data file and identified as uncodeable in variable Z3CODABL.

[98] There were six nonconsecutive coding weeks where coders challenged the reliability score and changes in scoring resulted because the coder pointed out a discrete behavior that the standard coders had missed. However, it was often the case that other coders who had already been declared reliable for the week had also missed this behavior. It was decided, in these instances, to allow all coders to code, since aside from that one code they were in 100 percent agreement (i.e., within two codes) for each column on the coding sheet.

[99] This circumstance occurred once. Reliability and retraining protocols were followed.

Coders also noted challenges encountered when scoring tapes. For example, if the interviewer signaled to the dyad to stop reading and move on to the second bag (which was a violation of the administration protocol), it was noted. Other challenges encountered by coders included the following: the mother or child, or both, lost interest in the book; the dyad was difficult to hear or understand; there was interference from others during the activity; book reading went longer than the allowable time; there were problems with the equipment; the dyad did not open the bags in the correct order; the child would not let the mother see the book; or the child was listening to the mother read *Corduroy* but was skimming through another book at the same time. Variable Z3CODDIF can be used to identify the coding challenges for individual cases on the data file.

6.3.4.2 Reserve Codes

Uncodeable tapes were assigned reserve codes, though the specific reserve code used depends on the type of variable. For discrete RAPT behaviors and the quality indicators, uncodeable tapes received -1 (Not applicable). For the ancillary information (e.g., dyad language, language of the book), uncodeable tapes received -9 (Not ascertained).

Additionally, reserve codes were used to indicate situations in which not all sections (i.e., prereading, during reading, and postreading) could be coded. A small number of cases (fewer than 50) received -9 (Not ascertained) codes on prereading because the DVD recording began after the interaction had started. Consequently, the coder could not ascertain whether any prereading occurred. Likewise, if a dyad lost interest in the book, the remainder of the book reading activity was coded -9 (Not ascertained). A total of about 100 cases have -9 codes on the postreading variables for a variety of reasons, including the dyad abandoning the book reading task before finishing the book and equipment difficulties rendering the postreading segment uncodeable (e.g., the DVD recorder cut off). Other reasons for a Not ascertained code on postreading variables include there being no clear end to the reading activity at the end of the task time and interference issues or problems with the administration protocol.

6.3.4.3 Second Read Flag

A small number of dyads (fewer than 50) chose to read the book a second time. Rather than include a set of variables pertaining to the second read for these cases and code the remaining cases as -1 (Not applicable), a flag variable (Z3FLG2RD) was created to identify the dyads that read the book a second time. Note that behaviors observed during the first read were collapsed with behaviors observed during the second read.[100] Thus, a user can identify dyads that chose to read the book again but cannot analyze how behaviors from the first read may have been different from behaviors evidenced during the second read.

[100] If the dyad completed the story and then opened the book and began rereading the story from the beginning, a second reading was noted. The behaviors coded during the second reading were identical to those coded during the first reading. If a during-reading behavior occurred at all, either during the first reading, the second reading, or both, it was coded as present.

6.3.4.4 Language of Interaction

The language of the interaction for the RAPT sample (identified in variable Z3PRTLNG) was determined by viewing the book reading section of the Two Bags Task interaction. In most cases (about 650), the dyad spoke English throughout the Two Bags Task interaction and read *Corduroy* in English. In about 50 cases, the dyad read from a Spanish language book (a Spanish translation of *Corduroy*), and a coder fluent in Spanish scored the interaction. Tapes not in English or Spanish were not able to be scored. However, in some cases (less than 50) the dyad was bilingual and conducted the interaction in both English and their home language. Z3PRTLNG notes when this occurred. In contrast, the Two Bags Task language variable X3TBLNG identifies only one language, which results in some inconsistency between Z3PRTLNG and X3TBLNG. Trained coders working on the Two Bags Task (who were different than the RAPT coders) assigned a language to the case based on whether the English-only coder felt that he or she could score the tape; if there was sufficient interaction in another language, the case was assigned either to a bilingual Spanish coder or (if not in Spanish) rendered uncodeable, and X3TBLNG was set to the non-English language. In a few cases, the dyad spoke one language (e.g., English) during the reading portion and then switched to another language (e.g., their home language) for the Play-Doh® portion of the Two Bags Task. Consequently, Z3PRTLNG does not match X3TBLNG in these cases.

Chapter 7
Indirect Child Assessments of Socioemotional Skills and Behaviors

The ECLS-B included a number of indirect child assessments administered through the parent interview, early care and education provider (ECEP) interview, and Teacher Self-Administered Questionnaire (TSAQ) to supplement those collected in the direct child assessment.[101] Section 7.1 discusses the development of the parent-reported language items used during the preschool wave of data collection and provides basic descriptive statistics on item performance. Section 7.2 discusses the development of socioemotional items used during the preschool, kindergarten 2006, and kindergarten 2007 waves of data collection. Descriptive statistics also are presented on item performance.

7.1 Parent Report of Children's Language Development

Children's vocabulary growth is generally regarded as an important predictor of school readiness and achievement. During the preschool wave, the assessment of children's vocabulary using items from the Peabody Picture Vocabulary Test (PPVT) on the child direct cognitive assessment was augmented by parent reports of child vocabulary using a standard word list (similar in format to the 2-year wave but updated in content for preschool-aged children) and general language skills.

The primary parent-reported measure of children's vocabulary is derived from the MacArthur Communicative Development Inventory (M-CDI) (Fenson et al. 1994), which is widely used in child development research. The M-CDI has an infant form for children from 8 to 16 months and a toddler form for children from 16 to 30 months of age. The toddler form includes a word list of 680 words in 22 semantic categories plus another 125 items that assess morphological and syntactic development. Scores are obtained by summing the words the child can say. Age-appropriate norms are available for the M-CDI.

Because the M-CDI is too long for survey administration and is designed for use only with children from 16 to 30 months of age, the M-CDI authors were consulted in the development of a shorter measure for use in the preschool wave. Accordingly, the authors created a vocabulary list to be more appropriate for preschool-aged children and suitable for fielding as part of the ECLS-B. This new version of the M-CDI is referred to as the M-CDI-IV and differs somewhat in content from the M-CDI for toddlers, although its form and function are basically the same. The individual items are included in the data file so that users may evaluate them and combine them as they wish. The M-CDI-IV is expected to measure children's language

[101] No indirect measures of the child were included in the wrap-around early care and education provider (WECEP) interview. However, all children eligible for a WECEP interview, except homeschoolers, also were eligible for a TSAQ; therefore, indirect child assessments of socioemotional skills were designed to be administered to the majority of sample children.

as well as the M-CDI toddler version, but it has not been extensively evaluated. Table 61 provides the 25 words that were included, the variable names and variable labels, and the weighted percentage of ECLS-B children whose parent reported the child could say the target word when they were preschool aged.[102] Parent respondents also rated their children's general communication skills using items taken from a measure of expressive and receptive language developed by Leventhal (1998). Table 62 summarizes these six items and provides the weighted percentage of ECLS-B preschool-aged children who displayed these language skills with varying degrees of frequency. The analyst is cautioned that the ECLS-B preschool wave word list and supplementary questions should not be considered the equivalent of the M-CDI and that the M-CDI norms do not apply. The user is encouraged to examine the factor structure of this set of items.

[102] Note that because these items were administered as part of the parent interview using audio computer-assisted self-interviewing, the items were not administered to parents who did not speak either English or Spanish.

Table 61. Preschool wave child vocabulary items in the Parent CAPI Instrument: 2005–06

Does child say…?	Variable name	Weighted percent for "yes" values[1]
Hungry	P3SYHNGR	97.7
Baby	P3SYBABY	98.2
Doctor	P3SYDCTR	97.1
Down	P3SYDOWN	98.3
Bird	P3SYBIRD	98.1
Fruit	P3SYFRT	93.9
Triangle	P3SYTRI	86.4
Turtle	P3SYTRTL	94.4
Plant	P3SYPLNT	90.8
Last	P3SYLAST	90.4
Caterpillar	P3SYCTRP	81.7
Castle	P3SYCSTL	87.5
Excited	P3SYEXCT	75.9
Stamp	P3SYSTMP	76.5
Parent	P3SYPRNT	69.8
Lucky	P3SYLCKY	71.7
Furniture	P3SYFRNT	73.1
Drip	P3SYDRIP	66.0
Measure	P3SYMSR	56.1
Calm	P3SYCALM	53.6
Lonely	P3SYLNLY	52.8
Dive	P3SYDIVE	49.5
Skeleton	P3SYSKLT	51.2
Uncomfortable	P3SYUNCM	42.9
Courage	P3SYCRG	29.8

[1]CAPI skip rules prevented parents from continuing with more difficult items when three previous items had been answered "no." Therefore, some children were skipped out of more difficult words. To create estimates for this table, children who were skipped out of more difficult words were analyzed as not being able to say the word.

NOTE: The audio computer-assisted self-interviewing portion of the interview was completed only in English or Spanish; if the parent spoke another language to conduct the interview, these items were skipped. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data. Only those cases with a valid weight are included in the table (n = approximately 8,400). CAPI = computer-assisted personal interview.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Table 62. Preschool wave child language use items in the Parent CAPI Instrument: 2005–06

Variable name	Item description	Response option	*n*	Weighted percent
P3SPKCLR	Child speaks clearly so a stranger can understand	(1) Always	4,450	53.0
		(2) Frequently	2,500	27.3
		(3) Sometimes	1,600	15.9
		(4) Rarely	250	2.9
		(5) Not at all	100	1.0
P3REFERS	Child refers to self as "I"	(1) Always	5,200	61.3
		(2) Frequently	1,600	18.0
		(3) Sometimes	1,300	13.1
		(4) Rarely	400	4.1
		(5) Not at all	350	3.6
P3ATTEN	Child is able to get attention of listener	(1) Always	6,250	74.4
		(2) Frequently	1,900	19.3
		(3) Sometimes	700	5.6
		(4) Rarely	50	0.6
		(5) Not at all	50	0.2
P3GREETS	Child uses appropriate social greetings	(1) Always	3,950	46.9
		(2) Frequently	2,900	32.4
		(3) Sometimes	1,750	17.7
		(4) Rarely	200	2.1
		(5) Not at all	100	0.8
P3LISTNR	Child is a good listener	(1) Always	2,100	23.7
		(2) Frequently	3,300	38.5
		(3) Sometimes	3,150	34.6
		(4) Rarely	300	2.7
		(5) Not at all	50	0.5
P3WAITS	Child waits his or her turn to speak	(1) Always	500	5.6
		(2) Frequently	2,050	23.5
		(3) Sometimes	4,800	54.0
		(4) Rarely	1,200	13.7
		(5) Not at all	300	3.2

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data. Only those cases with a valid weight are included in the table. CAPI = computer-assisted personal interview.

Rounds to zero.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

7.2 Socioemotional Skills and Behaviors

The ECLS-B collected information on children's socioemotional development from a number of sources, specifically parents, early care and education providers, and teachers. At the preschool wave, socioemotional items were included in the parent and ECEP interviews. At kindergarten 2006, the items were asked again of parents and providers, as well as teachers via the teacher questionnaire introduced with that wave. At kindergarten 2007, socioemotional information was collected in the parent and teacher instruments only, because the ECEP was not administered.

Early in the ECLS-B design phase, key socioemotional constructs related to preschool children's early learning experiences were identified for possible inclusion in the study. The key

constructs identified were prosocial skills, approaches toward learning, problem behaviors and emotions, emotional knowledge, temperament, and friendship. The Preschool and Kindergarten Behavior Scales (PKBS-2; Merrell 2003) was identified as a potential instrument to assess the socioemotional constructs of interest. The PKBS-2 is recognized as a strong measure of prosocial skills and also contains many items that address other constructs of interest in the ECLS-B, such as problem behaviors and emotional knowledge.

Because the full PKBS-2 was too long for administration in the ECLS-B, a subset of items was chosen for use in the study. Items were selected according to (1) their high item-to-total (subtest) correlations and (2) their relevance to the constructs of interest. Some items were modified on the basis of expert review. Some Social Skills Rating System (SSRS) (Gresham and Elliott 1990) socioemotional items also were included in support of the items selected from the PKBS-2. Finally, some items were created specifically for the ECLS-B to provide more comprehensive measurement of some of the constructs listed above.

Items from the social skills scale of the PKBS-2 measured prosocial behavior (empathy, cooperation, friendliness), friendship (interactions and involvement with other children), and emotional knowledge (understanding of emotions). Items were selected from the problem behavior scale to measure internalizing problems (emotions and related behaviors within the child that impede social interaction), externalizing problems (overt and aggressive actions), and temperament (individual differences in arousal and emotions, including attention span and inhibition). The PKBS-2 did not have items measuring one of the ECLS-B constructs of interest, approaches toward learning (tendencies, behaviors, and skills that support a positive attitude about learning). Items for measuring this construct were taken from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), which had been adapted from the SSRS.

Pilot testing was conducted on approximately 50 PKBS-2, SSRS, and newly developed items both to assess how these items are understood by respondents and to time their administration. Twelve parent interviews and 12 early care and education provider interviews were conducted on a sample recruited from child care centers.[103] For each interview type, 10 of the interviews were conducted using cognitive interviewing techniques and two were done without these techniques to most closely mimic actual data collection conditions. All participants responded to the full set of items. The completion time for the entire set of items tested was about 20 minutes for parents and 15 minutes for caregivers (who were not administered the friendship items).

Prior to the national data collection, wording modifications were made to three of the items on the basis of cognitive interview results. For example, “It’s her turn!” was added to

[103] Pilot testing of the indirect socioemotional items took place prior to the preschool wave (i.e., before the teacher instrument that would be implemented for the kindergarten 2006 wave had been developed). The items were not tested on teacher respondents.

"Stands up for other children's rights ('That's his!' or 'It's her turn!')" to reflect turn-taking situations. Also, the response scale was changed for all items (except for the two friendship items that were in a different format) from a 4-point to a 5-point scale by adding "very often" (i.e., never, rarely, sometimes, often, very often). Additionally, it was decided that children's prosocial skills, problem behaviors and emotions, emotional knowledge, temperament, and approaches to learning would be assessed together in one series of items: the socioemotional battery. Items included to measure friendship were not grouped with the above items but rather asked separately because they had a different response option format. The specific items included in the instruments varied slightly by instrument, as shown in table 63.

7.2.1 Parent Report

Table 64 provides information on the frequency distributions for the socioemotional items asked in the parent interview in the preschool, kindergarten 2006, and kindergarten 2007 collections and notes the sources of the items. Analysts may want to conduct factor analyses to explore the possibility of combining items to represent a particular construct of socioemotional development using these indirect measures.

Table 63. Socioemotional items by instrument: preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

Variable description	Parent variable name[1]	ECEP variable name	Teacher variable name
CHILD ACCEPTED BY OTHER CHILDREN[2]	P*LIKED	J*ACCPTD	T*ACCPTD
CHILD MAKES FRIENDS EASILY[2]	P3MKFRND	J*MKFRND	T*MKFRND
CHILD SHOWS EAGERNESS TO LEARN[3]	P*EAGER	J*EAGER	T*EAGER
CHILD PAYS ATTENTION WELL[3]	P*PAYATT	J*PAYATT	T*PAYATT
CHILD WORKS/PLAYS INDEPENDENTLY[3]	P*NDEPND	J*NDEPND	T*NDEPND
CHILD KEEPS WORKING UNTIL FINISHED[3]	P*FINISH	J*FINISH	T*FINISH
CHILD SHARES BELONGINGS WITH OTHERS[2]	P3SHARES	J*SHARES	T*SHARES
CHILD STANDS UP FOR OTHERS' RIGHTS[2]	P*STNDUP	J*STNDUP	T*STNDUP
CHILD COMFORTS OTHERS[2]	P*COMFRT	J*COMOTH	T*COMOTH
CHILD TRIES TO UNDERSTAND OTHERS[2]	P*UNDRST	J*UNDRST	T*UNDRST
CHILD ACTS IMPULSIVELY[2]	P*MPULSV	J*MPULSV	T*MPULSV
CHILD DISRUPTS OTHERS[2]	†	J*DISRPT	T*DISRPT
CHILD IS OVERLY ACTIVE[2]	P*ACTIVE	J*ACTIVE	T*ACTIVE
CHILD HAS DIFFICULTY CONCENTRATING[2]	P3CONCEN	J*CONCEN	T*CONCEN
CHILD IS RESTLESS/FIDGETY[2]	†	J*FIDGET	T*FIDGET
CHILD HAS TEMPER TANTRUMS[2]	P*TEMPER	J*TEMPER	T*TEMPER
CHILD IS PHYSICALLY AGGRESSIVE[2]	P*AGRESS	J*AGRESS	T*AGRESS
CHILD ANNOYS OTHER CHILDREN[2]	P*ANNOYS	J*ANNOYS	T*ANNOYS
CHILD SEEMS UNHAPPY[2]	P*UNHAPY	J*UNHAPY	T*UNHAPY
CHILD WORRIES ABOUT THINGS[3]	P*WORRY	J*WORRY	T*WORRY
CHILD ACTS SHY[2]	†	J4ACTSHY	T*ACTSHY
CHILD SHOWS IMAGINATION[4]	P4IMAGNE P5IMAGNE	J4SHWIMG	T*SHWIMG
CHILD INVITED TO PLAY BY OTHER CHILDREN[2]	P*INPLY	†	†
CHILD VOLUNTEERS TO HELP OTHERS[4]	P*VOLNTR	†	†
CHILD ACCEPTS IDEAS[3]	P4ACCEPT P5ACCEPT	†	†
CHILD USES WORDS TO DESCRIBE FEELINGS[5]	P*USWRDS	†	†
CHILD - COPYRIGHTED - ANGRY	P*ANGRY	†	†
CHILD INVITES OTHER CHILDREN TO PLAY[2]	P*INVITE	†	†
CHILD ADJUSTS TO NEW SITUATIONS[3]	P4ADJUST P5ADJUST	†	†
CHILD DESTROYS OTHERS' THINGS[2]	P*DESTRY	†	†
CHILD TRIES NEW THINGS[2]	P4TRYNEW P5TRYNEW	†	†

† Not applicable; question from which this variable is derived was not asked in this instrument.

[1] If the variable name has an asterisk in the second position (e.g., P*INPLY), the variable was asked at multiple data collection waves (preschool, kindergarten 2006, and kindergarten 2007). To determine the variable name for a particular wave, substitute the asterisk with a 3 (preschool), 4 (kindergarten 2006), or 5 (kindergarten 2007). For example, P3INPLY is the preschool variable name and P4INPLY is the equivalent kindergarten 2006 variable name. The parent instrument was used at all three data waves, while the early care and education provider interview was fielded only in the preschool and kindergarten 2006 waves, and the teacher survey was fielded in the kindergarten waves (2006 and 2007) only.

[2] Preschool and Kindergarten Behavior Scales–Second Edition (PKBS-2) item.

[3] Social Skills Rating System (SSRS) item.

[4] Family and Child Experiences Study (FACES) item.

[5] Item developed new for the ECLS-B.

NOTE: The PKBS-2 and SSRS are copyrighted materials. These assessments may be requested from the National Center for Education Statistics once publisher permission has been obtained. See "Guidelines for the Release and Use of ECLS-B Copyrighted Measures" at http://nces.ed.gov/ecls/pdf/Birth/ECLSB_Copyright_Guidelines.pdf.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Table 64. Item frequency distributions for parental report of children's socioemotional skills and behaviors, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08

			Preschool		Kindergarten 2006		Kindergarten 2007	
Variable name[1]	Variable label	Response option	*n*	Weighted percent	*n*	Weighted percent	*n*	Weighted percent
P*INPLY[2]	CHILD INVITED TO PLAY BY OTHER CHILDREN	(1) Never	350	3.3	200	2.9	50	3.0
		(2) Rarely	500	5.0	350	4.3	100	3.4
		(3) Sometimes	1,900	19.9	1,650	21.8	400	20.2
		(4) Often	3,600	40.8	2,750	39.7	800	40.4
		(5) Very often	2,550	30.9	2,000	31.2	550	33.0
P*EAGER[3]	CHILD SHOWS EAGERNESS TO LEARN	(1) Never	100	0.7	50	0.4	#	0.7
		(2) Rarely	150	1.5	100	1.0	#	0.7
		(3) Sometimes	1,000	9.9	800	10.3	200	8.7
		(4) Often	3,300	35.4	2,550	35.0	700	36.7
		(5) Very often	4,400	52.4	3,500	53.3	950	53.1
P*VOLNTR[4]	CHILD VOLUNTEERS TO HELP OTHERS	(1) Never	300	3.2	200	3.1	50	3.0
		(2) Rarely	600	6.4	400	5.8	100	4.1
		(3) Sometimes	2,750	29.5	2,000	27.2	500	26.6
		(4) Often	3,150	35.7	2,450	36.1	700	35.4
		(5) Very often	2,050	25.2	1,850	27.8	550	30.9
P*LIKED[2]	CHILD IS LIKED BY OTHERS	(1) Never	50	0.4	#	0.1	#	0.1
		(2) Rarely	100	0.8	50	0.7	#	0.9
		(3) Sometimes	750	7.3	550	6.6	150	6.3
		(4) Often	3,600	39.0	2,850	39.1	750	38.7
		(5) Very often	4,400	52.5	3,550	53.6	950	54.0
P3SHARES[2]	CHILD SHARES BELONGINGS	(1) Never	100	0.9	†	†	†	†
		(2) Rarely	400	3.8	†	†	†	†
		(3) Sometimes	3,200	35.0	†	†	†	†
		(4) Often	3,550	40.8	†	†	†	†
		(5) Very often	1,650	19.5	†	†	†	†
P4ACCEPT[3]	CHILD ACCEPTS IDEAS	(1) Never	†	†	50	0.5	#	0.6
P5ACCEPT		(2) Rarely	†	†	250	3.8	50	3.6
		(3) Sometimes	†	†	2,200	30.6	600	30.1
		(4) Often	†	†	2,950	43.5	850	43.6
		(5) Very often	†	†	1,550	21.5	400	22.1

See notes at end of table.

Table 64. Item frequency distributions for parental report of children's socioemotional skills and behaviors, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08—Continued

			Preschool		Kindergarten 2006		Kindergarten 2007	
Variable name[1]	Variable label	Response option	*n*	Weighted percent	*n*	Weighted percent	*n*	Weighted percent
P*AGRESS[2]	CHILD IS PHYSICALLY AGGRESSIVE	(1) Never	1,950	22.5	2,100	30.6	600	34.1
		(2) Rarely	3,650	42.7	2,800	41.9	800	40.5
		(3) Sometimes	2,550	27.8	1,600	21.4	400	19.5
		(4) Often	500	5.0	300	4.6	50	3.8
		(5) Very often	200	2.0	150	1.5	50	2.1
P*UNHAPY[2]	CHILD SEEMS UNHAPPY	(1) Never	3,450	40.9	2,550	39.0	700	38.3
		(2) Rarely	4,150	46.0	3,350	47.9	900	46.5
		(3) Sometimes	1,150	11.6	1,000	12.2	300	13.4
		(4) Often	100	1.0	50	0.8	#	1.5
		(5) Very often	50	0.5	#	0.1	#	0.3
P*COMFRT[2]	CHILD COMFORTS OTHER CHILDREN	(1) Never	250	2.5	200	2.4	50	2.4
		(2) Rarely	550	5.8	450	5.7	100	4.3
		(3) Sometimes	3,150	33.9	2,500	34.9	650	33.4
		(4) Often	3,250	38.6	2,450	37.4	700	39.4
		(5) Very often	1,600	19.3	1,300	19.6	350	20.5
P*USWRDS[5]	CHILD USES WORDS TO DESCRIBE FEELINGS	(1) Never	250	2.3	100	1.0	50	1.1
		(2) Rarely	500	4.7	250	3.5	50	3.6
		(3) Sometimes	2,200	22.6	1,400	17.7	400	18.7
		(4) Often	3,300	37.3	2,650	39.0	700	35.4
		(5) Very often	2,700	33.2	2,550	38.8	700	41.3
P*ANGRY[3]	CHILD - ANGRY	(1) Never	450	5.0	450	6.7	150	6.9
		(2) Rarely	2,700	29.7	2,250	31.7	700	37.0
		(3) Sometimes	3,800	43.7	2,900	42.3	800	39.5
		(4) Often	1,450	16.1	1,000	14.1	200	12.5
		(5) Very often	500	5.5	350	5.2	100	4.1
P*PAYATT[3]	CHILD PAYS ATTENTION WELL	(1) Never	50	0.7	50	0.5	#	0.5
		(2) Rarely	450	4.6	300	4.2	100	4.9
		(3) Sometimes	3,500	38.3	2,550	35.1	650	31.2
		(4) Often	3,550	41.0	2,850	43.2	800	42.9
		(5) Very often	1,300	15.4	1,250	17.0	350	20.5

See notes at end of table.

Table 64. Item frequency distributions for parental report of children's socioemotional skills and behaviors, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08—Continued

Variable name[1]	Variable label	Response option	Preschool		Kindergarten 2006		Kindergarten 2007	
			n	Weighted percent	*n*	Weighted percent	*n*	Weighted percent
P*NDEPND[3]	CHILD WORKS/PLAYS INDEPENDENTLY	(1) Never	100	0.6	50	0.5	#	0.6
		(2) Rarely	250	2.6	200	2.3	50	2.6
		(3) Sometimes	1,750	17.6	1,400	19.0	350	16.0
		(4) Often	3,850	43.6	3,000	43.1	800	41.5
		(5) Very often	2,900	35.6	2,300	35.2	650	39.3
P*MPULSV[2]	CHILD ACTS IMPULSIVELY	(1) Never	2,300	26.4	1,850	26.9	500	29.1
		(2) Rarely	3,200	36.4	2,650	39.3	800	38.7
		(3) Sometimes	2,400	26.8	1,850	25.8	450	23.7
		(4) Often	700	7.6	450	6.0	150	6.2
		(5) Very often	250	2.8	150	2.0	50	2.3
P*WORRY[3]	CHILD WORRIES ABOUT THINGS	(1) Never	2,100	22.9	1,300	17.9	350	16.7
		(2) Rarely	3,050	34.8	2,350	33.0	600	30.8
		(3) Sometimes	2,800	32.1	2,450	36.0	700	38.1
		(4) Often	700	7.7	650	9.8	200	11.8
		(5) Very often	200	2.5	200	3.3	50	2.7
P*ACTIVE[2]	CHILD IS OVERLY ACTIVE	(1) Never	1,100	11.1	1,000	12.3	250	13.3
		(2) Rarely	2,650	30.6	2,150	32.4	650	33.7
		(3) Sometimes	2,850	31.9	2,200	30.9	600	31.5
		(4) Often	1,350	14.9	950	14.3	250	12.3
		(5) Very often	1,000	11.5	700	10.1	150	9.3
P*INVITE[2]	CHILD INVITES OTHER CHILDREN TO PLAY	(1) Never	200	1.7	100	1.6	50	1.9
		(2) Rarely	350	3.3	200	2.3	50	2.5
		(3) Sometimes	1,750	17.7	1,350	16.8	350	17.3
		(4) Often	3,750	42.5	3,050	45.6	900	45.3
		(5) Very often	2,800	34.7	2,200	33.8	550	33.0
P*FINISH[3]	CHILD KEEPS WORKING UNTIL FINISHED	(1) Never	200	2.0	100	1.1	50	1.4
		(2) Rarely	1,000	11.0	600	7.7	150	5.9
		(3) Sometimes	4,050	45.4	2,850	41.8	750	38.7
		(4) Often	2,700	31.6	2,400	35.5	650	36.6
		(5) Very often	950	10.0	1,000	13.8	300	17.4

See notes at end of table.

Table 64. Item frequency distributions for parental report of children's socioemotional skills and behaviors, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08—Continued

			Preschool		Kindergarten 2006		Kindergarten 2007	
Variable name[1]	Variable label	Response option	*n*	Weighted percent	*n*	Weighted percent	*N*	Weighted percent
P*STNDUP[2]	CHILD STANDS UP FOR OTHERS' RIGHTS	(1) Never	300	2.7	200	1.9	50	1.9
		(2) Rarely	650	7.2	400	5.3	100	4.3
		(3) Sometimes	2,950	33.2	2,300	32.0	550	30.8
		(4) Often	3,250	37.5	2,650	38.9	800	38.9
		(5) Very often	1,650	19.3	1,400	21.9	400	24.0
P*TEMPER[2]	CHILD HAS TEMPER TANTRUMS	(1) Never	950	10.2	1,000	14.1	300	16.0
		(2) Rarely	2,850	33.2	2,450	36.1	700	37.7
		(3) Sometimes	3,550	40.8	2,600	36.8	650	33.8
		(4) Often	1,100	11.4	650	9.7	150	8.2
		(5) Very often	400	4.3	250	3.3	50	4.3
P4ADJUST[3] P5ADJUST	CHILD ADJUSTS TO NEW SITUATIONS	(1) Never	†	†	50	1.1	#	0.7
		(2) Rarely	†	†	450	5.6	100	5.6
		(3) Sometimes	†	†	2,250	30.9	600	28.5
		(4) Often	†	†	3,100	46.2	900	48.9
		(5) Very often	†	†	1,100	16.1	300	16.2
P3CONCEN[2]	CHILD HAS TROUBLED CONCENTRATING	(1) Never	1,150	13.2	†	†	†	†
		(2) Rarely	3,400	39.8	†	†	†	†
		(3) Sometimes	3,450	37.8	†	†	†	†
		(4) Often	700	7.1	†	†	†	†
		(5) Very often	200	2.0	†	†	†	†
P*ANNOYS[2]	CHILD ANNOYS OTHER CHILDREN	(1) Never	1,850	20.9	1,450	21.4	350	19.8
		(2) Rarely	4,000	46.1	3,100	45.0	850	47.1
		(3) Sometimes	2,450	27.6	1,950	27.6	550	27.1
		(4) Often	450	4.1	350	4.6	100	4.8
		(5) Very often	150	1.3	100	1.4	50	1.1
P*DESTRY[2]	CHILD DESTROYS OTHERS' THINGS	(1) Never	4,750	55.7	3,850	58.0	1,050	58.8
		(2) Rarely	2,900	32.0	2,250	31.4	600	29.8
		(3) Sometimes	950	9.7	650	8.3	200	9.1
		(4) Often	200	2.0	100	1.6	50	1.7
		(5) Very often	100	0.6	50	0.7	#	0.5

See notes at end of table.

Table 64. Item frequency distributions for parental report of children's socioemotional skills and behaviors, ECLS-B preschool, kindergarten 2006, and kindergarten 2007 data collections: 2005–06, 2006–07, and 2007–08—Continued

			Preschool		Kindergarten 2006		Kindergarten 2007	
Variable name[1]	Variable label	Response option	*n*	Weighted percent	*n*	Weighted percent	*N*	Weighted percent
P*UNDRST[2]	CHILD TRIES TO UNDERSTAND OTHERS	(1) Never	200	2.0	150	1.6	50	1.6
		(2) Rarely	600	6.2	400	5.3	100	5.1
		(3) Sometimes	3,200	34.7	2,500	35.3	650	34.1
		(4) Often	3,350	38.9	2,700	40.4	800	41.2
		(5) Very often	1,550	18.3	1,200	17.4	300	18.0
P3MKFRND[2]	CHILD MAKES FRIENDS EASILY	(1) Never	50	0.5	†	†	†	†
		(2) Rarely	250	2.2	†	†	†	†
		(3) Sometimes	1,150	10.8	†	†	†	†
		(4) Often	3,700	41.2	†	†	†	†
		(5) Very often	3,750	45.3	†	†	†	†
P4TRYNEW[2]	CHILD TRIES NEW THINGS	(1) Never	†	†	50	0.3	#	0.6
P5TRYNEW		(2) Rarely	†	†	200	2.4	50	1.3
		(3) Sometimes	†	†	1,400	18.3	400	19.0
		(4) Often	†	†	2,950	43.9	800	44.5
		(5) Very often	†	†	2,400	35.0	650	34.6
P4IMAGNE[4]	CHILD SHOWS IMAGINATION	(1) Never	†	†	50	0.3	#	0.3
P5IMAGNE		(2) Rarely	†	†	100	1.1	#	1.2
		(3) Sometimes	†	†	850	9.8	200	9.7
		(4) Often	†	†	2,700	37.9	750	38.4
		(5) Very often	†	†	3,300	50.9	900	50.5

† Not applicable; question from which this variable is derived was not asked in this wave.

Rounds to zero.

[1] If the variable name has an asterisk in the second position (e.g., P*INPLY), the variable was asked at all three data collection waves (preschool, kindergarten 2006, and kindergarten 2007). To determine the variable name for a particular wave, substitute the asterisk with a 3 (preschool), 4 (kindergarten 2006), or 5 (kindergarten 2007). For example, P3INPLY is the preschool variable name and P4INPLY is the equivalent kindergarten 2006 variable name.

[2] Preschool and Kindergarten Behavior Scales–Second Edition (PKBS-2) item.

[3] Social Skills Rating System (SSRS) item.

[4] Family and Child Experiences Study (FACES) item.

[5] Item developed new for the ECLS-B.

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. Estimates are weighted as follows: preschool estimates are weighted by W3R0, the weight appropriate for analyses of preschool parent and/or child assessment data; kindergarten 2006 estimates are weighted by W4R0, the weight appropriate for analyses of kindergarten 2006 parent and/or child assessment data; and kindergarten 2007 estimates are weighted by W5R0, the weight appropriate for analyses of kindergarten 2007 parent and/or child assessment data. Only those cases with a valid weight are included in the table. However, the cell counts are unweighted to demonstrate the distribution in each wave of the ECLS-B data collection. The PKBS-2 and SSRS are copyrighted materials. These assessments may be requested from the National Center for Education Statistics once publisher permission has been obtained. See "Guidelines for the Release and Use of ECLS-B Copyrighted Measures" at http://nces.ed.gov/ecls/pdf/Birth/ECLSB_Copyright_Guidelines.pdf.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

7.2.2 Early Care and Education Provider Report

Information about the socioemotional skills and behaviors of children not in kindergarten or higher who were receiving regularly scheduled nonparental care or education for at least 1 hour per week was collected through interviews with the child's ECEP during the preschool and kindergarten 2006 data waves. The ECEP interview was not fielded during the kindergarten 2007 data wave, because all the children were expected to be in kindergarten or higher. The majority of the socioemotional items included in the ECEP also were asked in the parent and teacher interviews. (See table 63 above.) Analysts may want to conduct factor analyses to explore the possibility of combining items to represent a particular construct of socioemotional development using these indirect measures from either the preschool or the kindergarten 2006 wave. Table 65 provides the frequency distributions for the ECEP items for each data wave.

Table 65. Item frequency distributions for early care and education provider report of children's socioemotional skills and behaviors, ECLS-B preschool (2005–06) and kindergarten 2006 (2006–07) data collections

			Preschool		Kindergarten 2006	
Variable name[1]	Variable label	Response option	*n*	Weighted percent	*n*	Weighted percent
J*ACCPTD[2]	CHILD ACCEPTED BY OTHER CHILDREN	(1) Never	#	0.1	#	#
		(2) Rarely	50	0.9	#	1.3
		(3) Sometimes	450	7.3	50	3.0
		(4) Often	2,350	39.9	450	35.1
		(5) Very often	2,900	51.8	750	60.6
J*MKFRND[2]	CHILD MAKES FRIENDS EASILY	(1) Never	50	0.4	#	0.7
		(2) Rarely	200	2.9	50	2.2
		(3) Sometimes	1,000	16.2	150	11.9
		(4) Often	2,300	39.1	500	39.2
		(5) Very often	2,250	41.4	600	46.0
J*EAGER[3]	CHILD SHOWS EAGERNESS TO LEARN	(1) Never	50	0.5	#	0.2
		(2) Rarely	250	3.6	50	1.8
		(3) Sometimes	1,100	18.6	200	13.2
		(4) Often	2,050	36.0	450	33.1
		(5) Very often	2,350	41.3	650	51.7
J*PAYATT[3]	CHILD PAYS ATTENTION WELL	(1) Never	50	0.6	#	0.7
		(2) Rarely	400	6.3	50	2.8
		(3) Sometimes	1,750	30.3	350	26.7
		(4) Often	2,150	37.9	500	36.8
		(5) Very often	1,400	24.9	400	33.1
J*NDEPND[3]	CHILD WORKS/PLAYS INDEPENDENTLY	(1) Never	50	0.9	#	1.0
		(2) Rarely	250	4.1	50	3.2
		(3) Sometimes	1,150	19.7	200	16.0
		(4) Often	2,500	44.4	500	37.2
		(5) Very often	1,800	30.8	500	42.6

See notes at end of table.

Table 65. Item frequency distributions for early care and education provider report of children's socioemotional skills and behaviors, ECLS-B preschool (2005–06) and kindergarten 2006 (2006–07) data collections—Continued

			Preschool		Kindergarten 2006	
Variable name[1]	Variable label	Response option	*n*	Weighted percent	*n*	Weighted percent
J*FINISH[3]	CHILD KEEPS WORKING UNTIL FINISHED	(1) Never	100	1.6	#	0.9
		(2) Rarely	450	7.3	50	4.4
		(3) Sometimes	1,600	28.2	300	23.5
		(4) Often	2,100	35.8	450	36.5
		(5) Very often	1,550	27.1	400	34.7
J*SHARES[2]	CHILD SHARES WITH OTHERS	(1) Never	50	0.8	#	0.9
		(2) Rarely	250	4.3	50	2.5
		(3) Sometimes	1,550	27.1	300	23.4
		(4) Often	2,500	43.0	550	43.4
		(5) Very often	1,400	24.9	400	29.8
J*STNDUP[4]	CHILD STANDS UP FOR OTHERS' RIGHTS	(1) Never	350	5.4	50	4.7
		(2) Rarely	800	12.8	100	8.6
		(3) Sometimes	1,850	32.2	400	29.6
		(4) Often	1,750	32.3	450	35.9
		(5) Very often	950	17.3	250	21.2
J*COMOTH[2]	CHILD COMFORTS OTHERS	(1) Never	350	5.3	50	5.8
		(2) Rarely	850	13.8	150	10.8
		(3) Sometimes	2,000	33.4	450	35.0
		(4) Often	1,650	31.2	400	29.0
		(5) Very often	900	16.4	250	19.4
J*UNDRST[2]	CHILD TRIES TO UNDERSTAND OTHERS	(1) Never	400	5.7	50	5.0
		(2) Rarely	850	13.5	150	10.0
		(3) Sometimes	2,050	35.7	500	39.3
		(4) Often	1,700	31.3	400	32.2
		(5) Very often	750	13.9	150	13.6
J*MPULSV[4]	CHILD ACTS IMPULSIVELY	(1) Never	1,750	31.0	450	39.6
		(2) Rarely	1,900	32.1	400	30.6
		(3) Sometimes	1,250	22.8	250	18.3
		(4) Often	600	9.9	100	7.5
		(5) Very often	250	4.2	50	4.0
J*DISRPT[2]	CHILD DISRUPTS OTHERS	(1) Never	1,600	27.6	400	32.0
		(2) Rarely	2,100	35.9	450	33.7
		(3) Sometimes	1,500	27.1	350	26.2
		(4) Often	400	6.9	100	6.4
		(5) Very often	150	2.6	#	1.7
J*ACTIVE[2]	CHILD IS OVERLY ACTIVE	(1) Never	1,700	28.1	450	37.3
		(2) Rarely	1,850	32.0	400	31.2
		(3) Sometimes	1,350	24.1	300	19.9
		(4) Often	600	10.5	100	8.7
		(5) Very often	300	5.4	50	2.9

See notes at end of table.

Table 65. Item frequency distributions for early care and education provider report of children's socioemotional skills and behaviors, ECLS-B preschool (2005–06) and kindergarten 2006 (2006–07) data collections—Continued

			Preschool		Kindergarten 2006	
Variable name[1]	Variable label	Response option	*n*	Weighted percent	*n*	Weighted percent
J*CONCEN[2]	CHILD HAS DIFFICULTY CONCENTRATING	(1) Never	1,300	22.7	350	27.9
		(2) Rarely	1,950	34.7	450	38.7
		(3) Sometimes	1,650	28.6	350	24.1
		(4) Often	600	10.1	100	7.3
		(5) Very often	250	3.9	50	2.0
J*FIDGET[2]	CHILD IS RESTLESS/ FIDGETY	(1) Never	1,450	25.0	400	33.5
		(2) Rarely	1,950	34.4	450	34.4
		(3) Sometimes	1,550	26.2	300	21.7
		(4) Often	600	10.2	100	7.7
		(5) Very often	250	4.2	50	2.6
J*TEMPER[2]	CHILD HAS TEMPER TANTRUMS	(1) Never	2,600	45.3	650	52.7
		(2) Rarely	1,600	26.1	350	24.3
		(3) Sometimes	1,100	20.5	200	15.8
		(4) Often	300	5.6	50	4.7
		(5) Very often	150	2.5	50	2.5
J*AGRESS[2]	CHILD IS PHYSICALLY AGGRESSIVE	(1) Never	2,900	50.3	700	57.8
		(2) Rarely	1,600	27.7	350	23.3
		(3) Sometimes	950	16.2	200	14.4
		(4) Often	250	4.1	50	3.2
		(5) Very often	100	1.7	#	1.3
J*ANNOYS[2]	CHILD ANNOYS OTHER CHILDREN	(1) Never	2,250	39.8	550	42.8
		(2) Rarely	1,950	34.6	450	31.2
		(3) Sometimes	1,200	20.2	250	22.1
		(4) Often	250	3.9	50	3.1
		(5) Very often	100	1.6	#	0.8
J*UNHAPY[4]	CHILD SEEMS UNHAPPY	(1) Never	2,550	45.9	600	49.7
		(2) Rarely	2,200	36.7	450	31.8
		(3) Sometimes	900	15.0	200	15.7
		(4) Often	100	1.8	50	2.6
		(5) Very often	50	0.6	#	0.3
J*WORRY[3]	CHILD WORRIES ABOUT THINGS	(1) Never	1,850	32.6	400	32.9
		(2) Rarely	1,950	33.2	450	33.4
		(3) Sometimes	1,550	27.8	350	27.3
		(4) Often	300	4.9	50	4.8
		(5) Very often	100	1.5	#	1.6
J4ACTSHY[2]	CHILD ACTS SHY	(1) Never	†	†	300	27.1
		(2) Rarely	†	†	350	23.6
		(3) Sometimes	†	†	500	37.8
		(4) Often	†	†	100	8.6
		(5) Very often	†	†	50	3.0

See notes at end of table.

Table 65. Item frequency distributions for early care and education provider report of children's socioemotional skills and behaviors, ECLS-B preschool (2005–06) and kindergarten 2006 (2006–07) data collections—Continued

			Preschool		Kindergarten 2006	
Variable name[1]	Variable label	Response option	*n*	Weighted percent	*n*	Weighted percent
J4SHWIMG[4]	CHILD SHOWS IMAGINATION	(1) Never	†	†	#	1.4
		(2) Rarely	†	†	50	1.9
		(3) Sometimes	†	†	300	19.9
		(4) Often	†	†	600	48.8
		(5) Very often	†	†	350	28.0

† Not applicable; variable not used in this wave.

Rounds to zero.

[1] If the variable name has an asterisk in the second position (e.g., J*ACCPTD), the variable was asked at both the preschool and kindergarten 2006 data collection waves. To determine the variable name for a particular wave, substitute the asterisk with a 3 (preschool) or 4 (kindergarten 2006). For example, J3ACCPTD is the preschool variable name and J4ACCPTD is the equivalent kindergarten 2006 variable name.

[2] Preschool and Kindergarten Behavior Scales–Second Edition (PKBS-2) item.

[3] Social Skills Rating System (SSRS) item.

[4] Family and Child Experiences Study (FACES) item.

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. The early care and education provider (ECEP) interview was not administered as part of the kindergarten 2007 wave. The ECEP weights, W33J0 (preschool) and W44J0 (kindergarten 2006), were used to obtain these statistics; only those cases with a valid weight are included in the table. However, the cell counts are unweighted to demonstrate the distribution in the ECLS-B preschool and kindergarten 2006 data collections. The PKBS-2 and SSRS are copyrighted materials. These assessments may be requested from the National Center for Education Statistics once publisher permission has been obtained. See "Guidelines for the Release and Use of ECLS-B Copyrighted Measures" at http://nces.ed.gov/ecls/pdf/Birth/ECLSB_Copyright_Guidelines.pdf.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06) and kindergarten 2006 (2006–07) data collections.

7.2.3 Teacher Report

Teachers of children enrolled in kindergarten or higher during the kindergarten 2006 and kindergarten 2007 data waves were asked a series of questions about the child's socioemotional skills and behaviors similar to the questions included in the parent and ECEP interviews. Additionally, the teacher questionnaire included several items that were asked in the ECLS-K teacher survey (see table 63 above). These additional items were not specifically pilot tested for the ECLS-B; they were thoroughly tested when used previously for the ECLS-K.

Table 66 provides the frequency distributions for the kindergarten 2006 and kindergarten 2007 collections and notes the sources for the teacher items. Analysts may want to conduct factor analyses to explore the possibility of combining items to represent a particular construct of socioemotional development using these indirect measures.

Table 66. Item frequency distributions for teacher report of children's socioemotional skills and behaviors, ECLS-B kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections

			Kindergarten 2006		Kindergarten 2007	
Variable name[1]	Variable label	Response option	*N*	Weighted percent	*n*	Weighted percent
T*ACCPTD[2]	CHILD ACCEPTED BY OTHER CHILDREN	(1) Never	#	0.1	#	#
		(2) Rarely	#	0.4	#	0.6
		(3) Sometimes	300	8.0	100	7.6
		(4) Often	1,500	38.6	550	35.8
		(5) Very often	2,000	52.9	800	56.0
T*MKFRND[2]	CHILD MAKES FRIENDS EASILY	(1) Never	#	0.4	#	0.6
		(2) Rarely	150	2.5	50	2.3
		(3) Sometimes	650	16.2	250	16.4
		(4) Often	1,500	40.4	500	34.7
		(5) Very often	1,450	40.5	600	46.1
T*EAGER[3]	CHILD SHOWS EAGERNESS TO LEARN	(1) Never	#	0.5	#	0.4
		(2) Rarely	200	4.8	50	2.8
		(3) Sometimes	800	20.0	250	18.3
		(4) Often	1,200	33.5	450	33.2
		(5) Very often	1,550	41.3	650	45.4
T*PAYATT[3]	CHILD PAYS ATTENTION WELL	(1) Never	50	1.0	#	1.2
		(2) Rarely	350	8.7	100	6.5
		(3) Sometimes	1,100	30.1	400	23.9
		(4) Often	1,200	31.3	450	34.0
		(5) Very often	1,100	28.9	450	34.3
T*NDEPND[3]	CHILD WORKS/PLAYS INDEPENDENTLY	(1) Never	50	0.5	#	0.8
		(2) Rarely	250	6.3	100	4.9
		(3) Sometimes	800	20.2	200	15.7
		(4) Often	1,400	37.7	500	33.8
		(5) Very often	1,350	35.4	600	44.8
T*FINISH[3]	CHILD KEEPS WORKING UNTIL FINISHED	(1) Never	50	1.3	#	1.4
		(2) Rarely	300	8.1	100	6.2
		(3) Sometimes	850	22.5	250	18.4
		(4) Often	1,250	32.8	450	28.8
		(5) Very often	1,350	35.3	600	45.4
T*SHARES[2]	CHILD SHARES WITH OTHERS	(1) Never	#	0.4	#	0.6
		(2) Rarely	100	2.0	50	2.3
		(3) Sometimes	550	14.7	200	13.8
		(4) Often	1,600	42.9	600	38.2
		(5) Very often	1,500	40.0	600	45.0
T*STNDUP[4]	CHILD STANDS UP FOR OTHERS' RIGHTS	(1) Never	200	4.0	100	4.2
		(2) Rarely	550	14.4	150	11.3
		(3) Sometimes	1,300	33.2	450	32.3
		(4) Often	1,150	31.9	450	32.0
		(5) Very often	600	16.6	250	20.2

See notes at end of table.

Table 66. Item frequency distributions for teacher report of children's socioemotional skills and behaviors, ECLS-B kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections—Continued

			Kindergarten 2006		Kindergarten 2007	
Variable name[1]	Variable label	Response option	*N*	Weighted percent	*N*	Weighted percent
T*COMOTH[2]	CHILD COMFORTS OTHERS	(1) Never	150	3.4	50	3.0
		(2) Rarely	600	15.1	200	12.2
		(3) Sometimes	1,400	37.6	500	36.1
		(4) Often	1,050	27.9	400	29.7
		(5) Very often	550	16.1	250	18.9
T*UNDRST[2]	CHILD TRIES TO UNDER-STAND OTHERS	(1) Never	200	3.9	50	3.7
		(2) Rarely	650	16.3	200	12.6
		(3) Sometimes	1,400	37.8	500	36.8
		(4) Often	1,050	28.6	400	30.7
		(5) Very often	500	13.3	200	16.3
T*MPULSV[4]	CHILD ACTS IMPULSIVELY	(1) Never	1,100	29.1	450	33.0
		(2) Rarely	1,400	38.7	500	38.1
		(3) Sometimes	750	20.5	250	18.7
		(4) Often	350	8.8	100	7.3
		(5) Very often	100	2.9	50	3.0
T*DISRPT[2]	CHILD DISRUPTS OTHERS	(1) Never	1,050	25.5	400	29.8
		(2) Rarely	1,500	41.4	600	41.5
		(3) Sometimes	850	22.6	300	20.2
		(4) Often	250	7.8	100	6.4
		(5) Very often	100	2.7	50	2.1
T*ACTIVE[2]	CHILD IS OVERLY ACTIVE	(1) Never	1,350	34.9	550	40.6
		(2) Rarely	1,300	35.0	500	33.4
		(3) Sometimes	650	17.8	250	16.2
		(4) Often	300	8.3	100	7.2
		(5) Very often	150	4.0	50	2.7
T*CONCEN[2]	CHILD HAS DIFFICULTY CONCENTRATING	(1) Never	850	22.4	350	27.3
		(2) Rarely	1,300	33.8	500	35.8
		(3) Sometimes	1,000	26.6	300	20.8
		(4) Often	450	11.4	150	11.0
		(5) Very often	200	5.7	100	5.0
T*FIDGET[2]	CHILD IS RESTLESS/ FIDGETY	(1) Never	1,200	31.3	500	36.4
		(2) Rarely	1,300	33.9	500	34.4
		(3) Sometimes	750	21.1	250	17.2
		(4) Often	350	8.9	150	8.6
		(5) Very often	150	4.8	50	3.5
T*TEMPER[2]	CHILD HAS TEMPER TANTRUMS	(1) Never	2,600	69.2	1,000	71.5
		(2) Rarely	700	18.1	300	17.6
		(3) Sometimes	300	8.2	100	7.2
		(4) Often	100	2.9	50	2.2
		(5) Very often	50	1.6	#	1.6

See notes at end of table.

Table 66. Item frequency distributions for teacher report of children's socioemotional skills and behaviors, ECLS-B kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections—Continued

			Kindergarten 2006		Kindergarten 2007	
Variable name[1]	Variable label	Response option	*N*	Weighted percent	*N*	Weighted percent
T*AGRESS[2]	CHILD IS PHYSICALLY AGGRESSIVE	(1) Never	2,500	65.6	950	67.7
		(2) Rarely	750	20.6	300	21.1
		(3) Sometimes	350	9.6	100	7.0
		(4) Often	100	3.2	50	2.7
		(5) Very often	50	1.0	#	1.5
T*ANNOYS[2]	CHILD ANNOYS OTHER CHILDREN	(1) Never	1,800	46.1	650	47.9
		(2) Rarely	1,100	31.3	450	31.8
		(3) Sometimes	600	16.4	250	14.4
		(4) Often	150	4.6	50	4.3
		(5) Very often	50	1.7	#	1.6
T*UNHAPY[2]	CHILD SEEMS UNHAPPY	(1) Never	1,550	40.8	550	40.7
		(2) Rarely	1,450	38.4	550	40.1
		(3) Sometimes	650	17.2	250	15.7
		(4) Often	100	2.8	50	2.6
		(5) Very often	50	0.7	#	0.9
T*WORRY[3]	CHILD WORRIES ABOUT THINGS	(1) Never	1,100	29.2	350	26.2
		(2) Rarely	1,450	37.6	550	39.0
		(3) Sometimes	1,050	27.1	400	28.4
		(4) Often	150	5.0	50	4.7
		(5) Very often	50	1.1	#	1.7
T*ACTSHY[2]	CHILD ACTS SHY	(1) Never	800	24.3	350	23.5
		(2) Rarely	1,100	28.4	450	32.0
		(3) Sometimes	1,300	33.8	450	31.2
		(4) Often	400	9.1	150	10.5
		(5) Very often	200	4.3	50	2.8
T*SHWIMG[4]	CHILD SHOWS IMAGINATION	(1) Never	50	1.6	50	2.1
		(2) Rarely	300	7.3	100	7.6
		(3) Sometimes	1,400	36.4	500	34.2
		(4) Often	1,500	40.7	600	40.1
		(5) Very often	500	14.0	200	16.1

Rounds to zero.

[1] To determine the variable name for a particular wave, substitute the asterisk with a 4 (kindergarten 2006) or 5 (kindergarten 2007). For example, T4ACCPTD is the kindergarten 2006 variable name and T5ACCPTD is the kindergarten 2007 variable name.

[2] Preschool and Kindergarten Behavior Scales–Second Edition (PKBS-2) item.

[3] Social Skills Rating System (SSRS) item.

[4] Family and Child Experiences Study (FACES) item.

NOTE: Sample sizes (*n*) have been rounded to the nearest 50. The teacher weights, W44T0 (kindergarten 2006) and W55T0 (kindergarten 2007), were used to obtain these statistics; only those cases with a valid weight are included in the table. However, the cell counts are unweighted to demonstrate the distribution in the ECLS-B kindergarten 2006 and kindergarten 2007 data collections. The PKBS-2 and SSRS are copyrighted materials. These assessments may be requested from the National Center for Education Statistics once publisher permission has been obtained. See "Guidelines for the Release and Use of ECLS-B Copyrighted Measures" at http://nces.ed.gov/ecls/pdf/Birth/ECLSB_Copyright_Guidelines.pdf.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 (2006–07) and kindergarten 2007 (2007–08) data collections.

Chapter 8
Coding of PreLAS and Fine Motor Items During the Preschool and Kindergarten Data Waves

As noted in previous chapters, several components of the direct child cognitive assessment required coding by centralized staff. These components included child responses to the PreLAS Let's Tell Stories items (see section 4.4) and several fine motor skills items (including copy forms and line through a curved path; see section 5.2.1). This chapter describes the coding operations for these items during the preschool (section 8.1), kindergarten 2006 (section 8.2), and kindergarten 2007 (section 8.3) data waves. The approach to assess and ensure reliability of coding was modified with each successive wave of data collection to produce the highest quality data for the study.

Throughout this chapter, we refer to "standard coders" and "field coders" in describing staff who completed the coding operations. Standard coders were RTI professional staff with experience in coding child-response assessment items for other early childhood studies. The standard coders received intensive direct training in applying each ECLS-B item's coding rules. They served as the standard for comparison in examining coding reliability for field coders during each wave.[104] Field coders were staff temporarily hired and trained by RTI to complete the majority of coding work at each wave. All standard and field coders were college graduates with fields of study in psychology, child development, early child education, sociology, or closely related fields. During each wave, coding operations were managed by a coding task leader,[105] with scientific oversight and training provided by the principal investigator.

This training of coders was guided by a coding manual developed specifically to support coding operations during each data collection wave. This manual drew from the source item scoring materials[106] and was constructed to provide procedural and technical details about coding operations for coders, including reliability procedures. The manual also included all scoring rules

[104] The process by which the standard coders provided comparison scores for reliability assessment took two forms during ECLS-B coding operations. With the "standard-coder" model used during the preschool wave, a sample of each field coder's scores was compared to those of one of the standard coders through a double-coding process. In the "standard-comparison" model (implemented during the course of kindergarten 2006 coding operations, and employed again during kindergarten 2007 coding), field coder scores were compared against scores developed by standard coders (and in some cases, the principal investigator) through consensus coding. The primary difference between these two methods is that the standard-coder model is a variant of an inter-rater model, where the standard coder is considered to be the gold standard in application of the coding scales used, while the standard-comparison model compares a field coder's score with a score arrived at through consensus coding by multiple expert coders (for the ECLS-B, the principal investigator and lead coder), a process that is expected to yield more accurate scores.

[105] During the preschool wave, one of the initial standard coders also served as the initial coding task leader.

[106] Source materials included the Early Screening Inventory-Revised [ESI-R] for copy forms items, the Bruininks-Oseretsky Test for the curved path item, and the holistic coding scheme provided in the PreLAS manual (Duncan and De Avila 1998) for the Let's Tell Stories items.

for the coded items, descriptions of protocols for maintaining data and materials security (e.g., access restrictions for the secure coding room), and the plan for initial and ongoing coder reliability assessment (e.g., criterion for passing certification during training).

During coding, a standardized scoring sheet was used to document problems encountered in the coders' review of the PreLAS and fine motor items in the Child Assessment Booklet. Feedback based on this sheet was forwarded to specific field interviewers on an as-needed basis[107] (e.g., if the field interviewer failed to record the child's response verbatim to a Let's Tell Stories item). In addition, periodic e-mails and newsletters about data quality (e.g., clarity of audiotape recordings) were sent to the entire field staff to reinforce the proper administration of all the PreLAS and fine motor items during each wave.

8.1 Coding Operations During the Preschool Data Wave

This section describes reliability procedures implemented for the Let's Tell Stories and fine motor items during the preschool wave, including the hiring and training of centralized coders and the methods used to assess reliability during the coding of audiotapes and information contained in the Child Assessment Booklets. During the preschool wave, a single set of coders was initially hired and trained to code these items as well as the Two Bags Task interactions, as described in section 6.2. The training of the preschool wave coders is described in section 8.1.1, with a discussion of certification procedures presented in section 8.1.2. The methods used to monitor reliability and results from the assessments of reliability conducted during coding operations are described in section 8.1.3.

8.1.1 Hiring and Training of Coders

Prior to the beginning of preschool wave coding, three standard coders were selected and trained by the RTI project director, with training guided by the coding manual developed for that wave. Standard coder training included a review of the scoring rubrics provided for each item by its original source assessment (i.e., the Early Screening Inventory-Revised [ESI-R] for copy forms items and the Bruininks-Oseretsky Test for the curved path item) and a series of scoring exercises to build consensus, or agreement on the scores assigned. The standard coders were trained to employ the holistic coding scheme provided in the PreLAS manual (Duncan and De Avila 1998) for the Let's Tell Stories items. This scoring used a 5-point scale, in addition to a "0" score applied for no response and for non-English responses to English stimuli. Copy forms items were scored 0 (fail) or 1 (pass); the line through a curved path item was scored fail (0), partial credit (1), or full credit (2). Training included direct instruction, practice scoring of cases from the preschool field test, and group coding and discussion of cases to arrive at common scores to be used as exemplars (i.e., consensus cases) during field coder training and certification. The coding manual also was revised, as needed, to clarify item scoring or coding procedures in response to questions that arose during this training.

[107] This was rarely necessary and only occurred one or two times during the course of data collection.

To conduct the primary coding operations for the preschool wave, a staff of 17 field coders was selected through an outside employment service. The field coders attended a 3-day training on the coding scales September 16–18, 2005. The training, which was held at an RTI facility in Research Triangle Park, NC, was planned and led by the RTI project director and coding task leader assigned to manage the coding work. The training covered the Let's Tell Stories, fine motor, and Two Bags Task items (see also section 6.2). The training included a review of coding procedures and policies; a review of the scoring guidelines for each item to be coded, using sample cases for each score point; and group and individual practice with cases drawn from the preschool field test.

8.1.2 Certification of Field Coders

Before field coders were allowed to code any data collected from the study children, they had to demonstrate that they could reliably code the data according to the scoring specifications outlined in the coding manual for both the PreLAS and fine motor items. This included the application of the holistic scoring scheme for the stories and the scoring rubrics for the individual copy forms and line through a curved path items. At the conclusion of training, field coders had to score a set of five cases previously coded by the coding task leader and standard coders.[108] Scoring was done using the instructions provided by the original instruments. Field coders were certified as reliable if they demonstrated 80 percent agreement with the previously assigned, or standard, scores. For the Let's Tell Stories items, percent agreement was calculated as the percentage of cases for which the field coder's scores were within 1 point of the standard score. To be certified to code the Let's Tell Stories items, each field coder had to demonstrate agreement on each of the two story items for at least four out of five cases in the certification set. For fine motor items, percent agreement was calculated as the percentage of cases for which the field coder's score exactly matched the standard score. For the fine motor items, field coders had to demonstrate reliability for each item on at least four out of the five cases in the certification set.

8.1.3 Reliability of Coding

A standard-coder method was used to assess the reliability of the PreLAS and fine motor items centrally coded at RTI during the preschool wave. With this approach, 5 percent of each coder's weekly production was double-coded by one of the three standard coders to assess reliability.[109] The double-coded cases (considered the reliability sample) were randomly selected for recoding by a standard coder after they had been scored by the field coders. Field coders did not know which cases would be selected for reliability assessment, and standard coders were blind to the scores assigned by the field coders. Disagreements in scoring were defined as the

[108] The coding task leader and standard coders created a series of certification cases for this purpose by independently coding the cases and then discussing and adjudicating differences together so that each case had a standard score assigned to it.
[109] One of these coders left RTI shortly after coding had begun. As a result, most of the double-coding was completed by two standard coders.

standard coder and field coder scores differing by more than a single point for the Let's Tell Stories items and differing by a single point or more for the fine motor items. Percent agreement for each coder was calculated for each item as the percentage of cases in each coder's 5 percent reliability sample that were in agreement as defined for the item type (fine motor item or story item). When a field coder's percent agreement in any item was below 80 percent, the standard coder reviewed with the field coder those cases for which the field coder's score was not in agreement with the standard coder's score (agreement defined as exact match for fine motor items, within 1 point for the story items).[110] Additionally, the next three to five cases[111] scored by the field coder whose percent agreement had been determined to be less than 80 percent were reviewed by a standard coder to ensure that the field coder was scoring the items appropriately.[112] Table 67 provides the results of the ongoing reliability assessment for the preschool Let's Tell Stories items, including the average percent agreement and the range of percent agreement, across all field coders, for each item. In all cases, irrespective of percent agreement score, the standard coder's scores were assigned to the double-coded case, replacing those of the field coder even when the percent agreement between scores was 100 percent.[113]

Table 67. Average ongoing reliability (percent agreement) for PreLAS 2000 Let's Tell Stories items, preschool data collection: 2005–06

Item	Number of reliability cases	Total number of cases coded	Average percent agreement	Range of percent agreement
Story 1	376	8,550	98.9	91.3–100.0
Story 2	376	8,550	98.1	90.0–100.0

NOTE: Total number of cases coded has been rounded to the nearest 50.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

Table 68 provides the results of the ongoing reliability assessment for the preschool fine motor items, including the average percent agreement and the range of percent agreement, across all field coders, for each item. Note that for the fine motor items, percent agreement was calculated as the percentage of reliability cases where field coder scores matched the scores given to each case by the standard coders.

[110] Because the field coders' percent agreement for the story items never fell below the 80 percent criterion, this procedure was not used for the story items during the preschool wave.
[111] The exact number of cases reviewed was at the discretion of the standard coder. The goal of this review was to identify the source of disagreement between the coder and the standard, so in instances where the error was easily identified and corrected, fewer cases would be reviewed.
[112] This procedure was modified during the kindergarten 2006 wave (as discussed in section 8.2) so that all of a field coder's cases coded during a period in which that coder's reliability fell below the criterion (other than cases for which a standard coder's score was available) were subsequently recoded.
[113] Because agreement was defined as being within 1 point, the standard coder's scores could be different from the field coder's scores even when percent agreement was 100 percent.

Table 68. Average ongoing reliability (percent agreement) for fine motor items used in the ECLS-B preschool data collection: 2005–06

Item	Number of reliability cases	Total number of cases coded	Average percent agreement	Range of percent agreement
Vertical line	378	8,750	90.2	82.3–100.0
Horizontal line	378	8,750	86.5	68.0–100.0
Circle	378	8,750	84.6	60.0–100.0
Addition sign/cross	378	8,750	88.9	70.6–100.0
Square	378	8,750	91.8	78.9–100.0
Triangle	378	8,750	93.9	88.2–100.0
Asterisk	378	8,750	92.6	76.5–100.0
Curved path	378	8,750	81.7	66.7–90.6

NOTE: Total number of cases coded has been rounded to the nearest 50.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool data collection, 2005–06.

8.2 Coding Operations During the Kindergarten 2006 Data Wave

The approach taken to monitor, maintain, and report reliability of coding of the Let's Tell Stories and fine motor items was changed for the kindergarten 2006 wave. The experience gained in the preschool wave Two Bags Task coding operation (see chapter 6) was used to inform the approach for coding the kindergarten 2006 PreLAS and fine motor data.[114] Some of the modifications were operational in that they related to methods for weekly reporting of reliability to NCES and so are not discussed in this report. Other modifications, however, concerned the role and timing of weekly reliability assessments. As described in the sections below, the approach was modified so that work completed by a field coder during a weekly period in which the coder was found to be unreliable (i.e., did not meet the reliability criteria) was subsequently recoded. Also, the reliability criterion for the PreLAS items was made more stringent, with percent agreement defined as exact agreement between the field coder and standard coder scores for all items (agreement for the fine motor items continued to be defined as exact match). The passing criteria also were set to 90 percent agreement for all items, thereby matching the reliability standards provided in the PreLAS manual.

This section details the reliability procedures employed for the kindergarten 2006 wave, including the hiring and training of coders and the assessment of reliability. One coder from the preschool wave of data collection was retained as the lead coder to allow for greater consistency across the preschool and kindergarten 2006 waves.[115] In addition, the initial method of reliability assessment (an adjudicated inter-rater model) that was implemented at the beginning of coding for the kindergarten 2006 wave was ultimately replaced during the kindergarten 2006 coding period by a standard-comparison model to increase efficiency during the coding operations. The

[114] The coding of preschool Two Bags Task interactions continued beyond the start of the kindergarten 2006 wave coding operations. Coding of preschool Let's Tell Stories and fine motor items had been completed by the time kindergarten 2006 coding began.
[115] This coder had been a field coder during the completion of the Let's Tell Stories and fine motor items for the preschool wave and was also a lead coder during Step 2 and Step 3 of the Two Bags Task coding (see section 6.2.3.2). The overall supervision of coding production continued to be provided by the coding task leader.

training of coders and the methods used to monitor reliability during coding operations—including the change in methods used during coding of kindergarten 2006 wave items—are discussed below.

8.2.1 Hiring and Training of Coders

Ten coders were selected through an outside employment service to attend training October 25–26, 2006, which incorporated instruction and certification on both the Let's Tell Stories items and the fine motor drawings. As in the preschool wave, all of the coders had a bachelor's degree in psychology, child development, early childhood education, sociology, or a related discipline, and preference was given to applicants with experience in data coding or child assessment. The training, which was held at an RTI facility in Research Triangle Park, NC, was planned and led by the RTI principal investigator with the assistance of the coding task leader assigned to manage the coding work. Before training, a coding manual similar in content to the one developed for the preschool wave was prepared. Instructions for scoring the Let's Tell Stories items were the same as those used during the preschool wave, as were the instructions for the fine motor items that also had been included in the preschool assessment. Instructions for scoring fine motor items introduced during the kindergarten 2006 wave were adapted from the ESI-R and included in the coding manual.

Prior to training, the principal investigator and lead coder, who had coded fine motor and Let's Tell Stories items during the preschool wave, created a set of reliability cases by selecting and independently scoring 40 audiotapes and Child Assessment Booklets from the kindergarten 2006 data collection. Scoring was done using the instructions provided for the original instruments. Immediately following completion of the independent coding, the principal investigator and lead coder discussed their scores and adjudicated differences to reach agreement. Through this consensus coding process, the principal investigator and lead coder established a set of standard scores for these cases. This process also allowed the lead coder's experience during the preschool wave to inform scoring rules for the kindergarten 2006 wave, with the goal of cross-wave consistency. The 40 reliability cases were used as training examples and practice cases and for certification activities. During training, the principal investigator provided an overview of the items to be scored. Exemplars from the above-described standard cases were provided for each item. Coders then had time during training to practice coding the items (individually and in groups) using cases drawn from the exemplars not selected for certification purposes.

8.2.2 Certification of Field Coders

To ensure coding reliability for both the PreLAS and fine motor items, a high criterion for initial coder reliability on an item-level basis was established. Field coders were required to demonstrate exact agreement on at least 90 percent of the items to be certified. They were certified on Let's Tell Stories using a set of 10 cases, each including 2 story items (20 items

total). For fine motor certification, field coders scored a set of 10 reliability cases for each of the five items (50 items total). These certification cases were drawn from the exemplars set aside for this purpose.

As in the preschool wave, field coders were trained to employ the holistic scoring criteria provided in the PreLAS manual for the Let's Tell Stories items. This scoring used a 5-point scale, in addition to a "0" score applied for no response and for non-English responses to English stimuli. Following the reliability procedures provided by the publisher for the PreLAS, coders had to achieve 90 percent agreement per story against the standard scores to be certified. When a coder demonstrated 90 percent reliability or higher on both stories,[116] the coder was considered certified and ready to code PreLAS stories during data collection.

Field coders were trained to score each fine motor item according to guidelines modified from the ESI-R and Early Childhood Longitudinal Study, Kindergarten Class of 1988–99 (ECLS-K) to be more applicable to the ECLS-B data; the copy forms items were scored 0 (fail) or 1 (pass). To be certified, coders had to achieve 90 percent agreement per item against the standard scores provided for these cases; that is, each coder's score had to agree with the standard scores for 9 out of 10 cases used as certification cases for each item or figure. When the coder demonstrated 90 percent reliability or higher on each of the fine motor items, the coder was considered certified and ready to code.

Field coders were allowed multiple attempts to be certified if they did not meet the criteria for certification on one or more of the individual items. In such cases, additional consensus-building instruction and practice were provided. These coders were then given a second set of certification items, drawn from additional Child Assessment Booklets and audiotapes set aside for this purpose. For example, if a field coder was certified on story 1 but not story 2, that coder received additional instruction on story 2 and then scored 10 new story 2 cases previously scored by the principal investigator and lead coder. Percent agreement was again calculated. A similar procedure was used for the fine motor items.

During the first kindergarten 2006 training session, field coders had difficulty achieving the criterion of 90 percent exact agreement for each of the fine motor and Let's Tell Stories items. Additional consensus-building instruction and practice were provided for coders who did not meet the criteria for certification on one or more fine motor items on the first attempt. During this first training period, some coders tried a second attempt at certification on the fine motor items for which they had not yet been certified. Other coders, however, preferred to first receive further instruction and to spend additional time scoring practice items before reattempting certification. As a result, an additional half day of training and consensus building for the fine

[116] The reliability standard for certification on the Let's Tell Stories items changed between the preschool and kindergarten 2006 waves. In the preschool wave, coders had to achieve 80 percent agreement within 1 point on each story to be certified. In accordance with the PreLAS manual, a more stringent standard was put in place for kindergarten 2006, requiring coders to achieve 90 percent exact agreement on each story to be certified. This stricter reliability standard for kindergarten 2006 has little implication for comparison with parallel items at preschool because, at the aggregate level, the difference is relatively small. See section 8.4 for further discussion of the implications of changes in reliability criteria across data waves.

motor items was held for all coders on October 31, 2006. This was followed by a second and, in some cases, third attempt to reach 90 percent agreement with the standard on items for which individual coders had not yet reached the certification criterion. As a result of these trainings, 10 coders were ultimately certified to code fine motor items.

An additional half day of training, practice, and consensus building on the PreLAS items was held on November 7, 2006, approximately 2 weeks after the initial training, to allow the field coders time to review additional home study materials developed for the PreLAS. The additional home study materials included further details about coding guidelines for the PreLAS and additional story examples for each score point. Following this training and consensus coding, all but one coder met the certification criterion (so, 9 of the 10 coders initially hired to code were certified for Let's Tell Stories). This coder did not code any Let's Tell Stories items for the ECLS-B.

8.2.3 Reliability of Coding Using Adjudicated Inter-Rater Methods

To ensure ongoing reliability, an adjudicated inter-rater reliability model (see Meisels et al. 1997) was initially used in the kindergarten 2006 wave.[117] This is not the same as the standard-coder model used during the preschool wave. At preschool, a standard coder recoded, or checked, 5 percent of each coder's work. In the kindergarten 2006 wave, a random sample of 20 percent of each coder's weekly cases was double-coded by another coder to establish inter-rater reliability. This new model, the adjudicated inter-rater reliability model, was designed to ensure that all the coders were coding in the same way, as a unified group, and that more of the coders' work could be assessed for reliability. The model offers efficiencies over the standard-coder model because the work load of the standard coder (in this model, the adjudicator) is reduced as reliability increases. Using this process, cases to be double-coded were distributed across all coders so that across the duration of the coding task, each coder's work would be double-coded by every other coder. A coder did not know which of his or her cases were chosen for double-coding, and the coder doing the double-coding did not have access to the first set of scores. After double-coding, discrepancies in the scores of the two coders were identified by the computer system used to manage coding operations. This system compared the scores of the two coders and identified cases to be adjudicated by the lead coder. Thus, this process was designed to provide both an ongoing estimate of coder agreement for the duration of coding and to provide for the opportunity to give feedback and, if necessary, corrective guidance to the coders, resulting in increased data quality.

The double-coding was completed within a targeted timeframe of 2 weeks following the first coding, and double-coded cases were prioritized by the computerized coding system to facilitate more rapid completion, enabling more timely assessment of reliability. For each of the

[117] As described in the text, however, while production was very high, reliability was initially low, especially for fine motor items, ultimately leading to the use of a standard-comparison model (see section 8.2.4) in which all coders were compared against a set of scores established by the standard coders for a set of standard cases. This transition allowed for more direct oversight of coders by the standard coder, resulting in increased reliability.

two story items, the average percent agreement for each coder was computed for all pairs of scores (i.e., each coder's score and the double coder's score). When a coder's average percent agreement was high (at least 90 percent on each story), the coder was considered reliable; when it was low, the coder was considered less reliable. An exception was made for coders with percent agreement on a single story during a single week between 85 and 89 percent.[118] Table 69 shows the average percent agreement and range of percent agreement for each of the Let's Tell Stories items, using the inter-rater percent agreement model. Note that overall the percent agreement was below the reliability criterion and only one coder met the criterion for each story.

Table 69. Inter-rater percent agreement for Let's Tell Stories items using an inter-rater reliability model, ECLS-B kindergarten 2006 data collection: 2006–07

Item	Number of coders	Number of reliability cases	Number of cases reviewed using this method	Average percent agreement	Percent of coders meeting reliability criteria[1]	Range of percent agreement[1]
Story 1	9	120	120	72.1	11.1	50–100
Story 2	9	120	120	72.7	11.1	50–100

[1] Coders who did not meet the 90 percent agreement criterion on a single story were deemed to be reliable if the percent agreement was at least 90 percent during the subsequent week of coding. As a result, for some coders, the percent agreement calculated for a given week could have been as low as 85 percent. Only scores provided by coders who met the reliability criteria during a given period were retained, as were scores for which two coders agreed and scores that were adjudicated (i.e., cases for which two coders' scores did not agree and were subsequently scored by the standard coder). The range indicated in this table reflects the work of all coders, including those determined to be nonreliable, and whose work was subsequently recoded.

NOTE: Reflects ongoing reliability of coders between December 1, 2006, and January 5, 2007.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 data collection, 2006–07.

For the fine motor items, a coder's work was considered to be reliable if the coder maintained a percent agreement of at least 85 percent on each figure each week, although provisions were made for coders with percent agreement of between 80 percent and 84 percent[119] on a single figure during a single week. The average percent agreement and range of percent agreement are given for each fine motor item in table 70.

[118] The criterion of 90 percent agreement was modified slightly when coders whose work was generally above the threshold fell below 90 percent on a single story. In these cases, if a coder's work on a single story was between 85 and 89 percent, but the percent agreement during the subsequent week was above 90 percent, the coder received some additional instruction, but the work was considered reliable. If the coder was below 90 percent for a second week, the coder received additional instruction on the scoring from the lead coder and had to be recertified on the story through a double-coding of the item or items for which the coder was determined to be nonreliable with the lead coder. Once recertified, the coder would continue work, and cases coded during the time period during which the coder was below 90 percent agreement were recoded. A coder who fell below 85 percent agreement on both stories received additional instruction and had to be recertified by the lead coder. Work completed by that coder during the period for which the coder was below the threshold was recoded by reliable coders.

[119] The criteria for fine motor items was adjusted to 85 percent agreement for each item because the greater number of fine motor items allowed for more tolerance at each item, than the PreLAS, which included only two items. This expectation was modified to consider coders whose work was generally above the threshold but below 85 percent on a single figure. In these cases, if a coder's work on a single figure was between 80 percent and 84 percent, but the percent agreement during the subsequent week was at least 85 percent, the coder's work was considered reliable for the week and the coder's scores remained on the file. The coder also received additional instruction on the scoring of the item(s) below the 85 percent criterion. If the coder's work was below 85 percent for a second week, the coder received additional instruction on scoring from the lead coder and had to be recertified on the figure through double-coding of 10 cases with the lead coder. Once recertified, the coder continued work, and cases coded during the time period in which the coder was below 85 percent agreement were recoded by reliable field coders. Coders who fell below 80 percent agreement on any figure received additional instruction and had to be recertified by the lead coder, and work completed by that coder during the period for which the coder was below 80 percent was recoded by reliable field coders.

Table 70. Inter-rater percent agreement for fine motor item coding using an inter-rater reliability model, ECLS-B kindergarten 2006 data collection: 2006–07

Item	Number of coders	Number of reliability cases	Average percent agreement	Percent of coders meeting reliability criteria	Range of percent agreement[1]
Square	10	1,001	89.6	100	83–96
Triangle	10	1,001	90.4	100	87–96
Asterisk	10	1,001	92.6	100	88–98
Circle-square	10	1,001	87.7	100	85–94

[1] Coders who did not meet the 85 percent agreement criterion on a single item were deemed to be reliable if the percent agreement was at least 80 percent for that week and was above 85 percent during the subsequent week of coding. As a result, for some coders, the percent agreement calculated for a figure during a given week could have been as low as 80 percent. The range of percent agreement includes all cases completed by coders.

NOTE: Data in this table are based on 10 coders who completed cases under the inter-rater reliability procedure between December 1, 2006, and January 5, 2007.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten data collection, 2006–07.

Using this approach, all scores for cases completed during periods for which a given coder was determined to be reliable were included in the final data file. Likewise, all scores for all cases in the reliability sample were included, either because the two coders' scores agreed or, in cases of disagreement, because the lead coder provided an adjudicated, final score. Cases completed by coders who did not meet the reliability criteria for a given week and that were not part of the reliability sample were recoded by coders who were shown to be reliable; ultimately this was done following a change in reliability approaches (see section 8.2.4). As a result, only scores that were provided by reliable coders or obtained through adjudication by the lead coder were included in the data file. Cases that were completed by coders shown to be nonreliable were recoded and their reliability confirmed before the scores were included.

8.2.4 Reliability of Coding Using Standard-Comparison Methods

The effectiveness of the inter-rater reliability method as a means of maintaining data quality was predicated on using coder agreement as a means of checking reliability with a timely adjudication process. The change from an inter-rater reliability assessment method to a standard-comparison approach was driven by the relatively low levels of reliability initially observed and for production reasons. The adjudicated inter-rater reliability approach assumes either high ongoing reliability (which was not seen on all items when coding production began) or timely adjudication and corrective feedback to address lapses in reliability. The fine motor coding operation progressed at a faster pace than scoring of the stories (nearly 4,300 fine motor cases had been scored by the time coders began coding PreLAS cases), and the adjudication workload for the lead coder was too much to manage well as a means of addressing coder reliability. On a weekly basis, the lead coder was expected to review all cases in the reliability sample for which the scores of the first and second coders disagreed, provide individual feedback to coders to prevent drift (variation in coders' scores over time), and provide clarification as unique or challenging situations arose. However, with all the tasks the lead coder needed to oversee, the adjudication could not occur and individualized feedback could not be delivered to coders in a

timely manner. As a result, the inter-rater reliability method could not be used to provide the best, most timely feedback to coders to maintain the high level of reliability desired, as evidenced by the data provided in table 69.

For these reasons, approximately 1 month into coding operations for the kindergarten 2006 wave a change was made to the method used to monitor ongoing reliability—the standard-comparison model was implemented.[120] This approach was modeled after the method used to assess reliability of the Two Bags Task coding (see section 6.2.3). It provides a standardized method of gauging ongoing reliability but is less time-consuming for the lead coder to monitor (see e.g., Biemer and Lyberg 2003). With this approach, coders completed a common set of cases (i.e., reliability cases) for each item each week for which standard scores[121] had been developed by the lead coder and principal investigator. The set included 10 stories (5 cases with 2 stories each) and 10 of each of the fine motor drawings. The cases were randomly selected from among cases not previously scored by coders. Coders were considered to be reliable if their percent agreement was 90 percent or greater on the set of 10 reliability items (10 stories and 10 of each fine motor item). Coders completed this set of reliability cases on their first day of coding each week and received immediate feedback on their performance. If a coder achieved 90 percent or greater agreement on the set of common cases, that coder continued scoring. If a coder did not achieve 90 percent agreement or greater, the coder was not allowed to code again until he or she had passed a second set of 10 cases for that failed item (e.g., a second set of 10 triangles).[122]

By the time the standard-comparison model was adopted for monitoring reliability, there had been some attrition in the field coding staff.[123] An attrition training session was conducted March 5–7, 2007, for staff hired to replace lost members of the coding staff who had been trained at the beginning of data collection. This training, which focused exclusively on the Let's Tell Stories items, was based on the initial training approach and materials. However, the content of the training program was enhanced based on the initial training experience to emphasize distinctions between specific scores and to provide more individual and group practice time prior to attempts at certification. Three field coders were trained and certified during this session; two of them ultimately joined the coding staff following certification (as per procedures described in

[120] Note that the standard-coder model used for these items during the preschool wave compared coder scores with those of a given standard coder. During the kindergarten waves, the standard-comparison method used scores from two experienced coders working each case to consensus as the comparison score.
[121] The standard scores were determined by consensus coding; the lead coder and principal investigator independently scored the tapes and adjudicated differences through discussion. In this way, a set of reliability tapes was created.
[122] This approach sets the reliability criteria higher than what was used under the adjudicated inter-rater reliability approach. Because this method checked reliability once (twice, for coders not initially meeting the criterion), rather than throughout each week of coding, the higher criteria allowed for some decrease in reliability during the week without going below the 85 percent agreement threshold used during the adjudicated inter-rater reliability model.
[123] Five of the original coders had left the project by this time, including one coder who had only certified initially on fine motor items, leaving five of the originally trained coders on the staff.

section 8.2.2).[124] As a result of this attrition and subsequent supplemental hiring and training, a total of seven field coders were available to complete work under this reliability model.

Once all coders completed the first set of reliability cases, all coding staff participated in a weekly staff meeting at which scores for this set of reliability cases were discussed, standard scores and rationale for each standard score were provided, and additional scoring guidance was given, if necessary. Following this meeting, coders who had met the 90 percent agreement criterion returned to production coding, while coders who had not reached that criterion were given the opportunity to receive more feedback and attempt a second set of reliability cases. Coders could not complete work on any new cases until they had demonstrated 90 percent agreement with the consensus scores. The performance of coders on the weekly reliability cases was generally quite strong. Percent agreement for coders using this comparison model is shown for the PreLAS items in table 71 and for the fine motor items in table 72.

Table 71. Coder percent agreement with standard scores for Let's Tell Stories items using a standard-comparison model, ECLS-B kindergarten 2006 data collection: 2006–07

Item	Number of coders	Number of reliability cases	Number of cases reviewed using this method	Average percent agreement	Percent of coders achieving agreement between 90 and 100 percent	Range of percent agreement
PreLAS stories	7	95	6,595	95.2	100	90–100

NOTE: Reflects ongoing reliability of coders, January 8, 2006, to May 11, 2007. Percent agreement data are based on each coder's attempts that met the percent agreement criteria; attempts that did not meet the criteria led to additional training and a second attempt. Range of percent agreement includes reliability sets through which coders were certified
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 data collection, 2006–07.

Table 72. Coder percent agreement with standard scores for fine motor item coding using a standard-comparison model, ECLS-B kindergarten 2006 data collection: 2006–07

Item	Number of coders	Number of reliability cases	Average percent agreement	Percent of coders meeting reliability criteria	Range of percent agreement
Square	7	190	96.5	100	90–100
Triangle	7	190	95.9	100	90–100
Asterisk	7	190	98.3	100	90–100
Circle-square	7	190	95.8	100	90–100

NOTE: Percent agreement data are based on each coder's attempts that met the percent agreement criteria; attempts that did not meet the criteria led to additional training and a second attempt. Range of percent agreement includes reliability sets through which coders were certified.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 data collection, 2006–07.

8.2.5 Source of Scores for the Kindergarten 2006 Wave

The transition from an inter-rater to a standard-comparison approach was intended to ensure reliability of coding while also maintaining operational efficiency. As shown in tables 70 and 72, the average percent agreement for fine motor items under either reliability method was above 85 percent, although the percent agreement against standard scores was generally at 95

[124] One of the coders trained withdrew from the project the week after certification.

percent or greater. This increase in percent agreement partially reflects a change in approach whereby a coder was not allowed to continue active coding until he or she had demonstrated at least 90 percent agreement against the standard scores each week. It should also be noted that the number of active coders decreased during the transition between reliability methods. Attrition is not unusual during coding operations stretching over several months, and the number of coders hired initially was higher than was necessary because some attrition during the coding process was expected.

The delayed start of story coding (due to the later date for additional training compared to completion of training for fine motor items) required further efficiency decisions to be made about the treatment of PreLAS cases initially coded under the inter-rater reliability method. PreLAS cases that had been double-coded as part of the inter-rater approach and for which the first and second coders' scores were in agreement were considered final. For cases that had been double-coded via the inter-rater reliability method but for which both coders' scores were not in agreement, the lead coder completed the adjudication process and provided feedback to coders on the results of this review. The scores assigned by the lead coder during this adjudication process also were considered final. However, all PreLAS cases that had not yet been checked for reliability under the adjudicated inter-rater reliability model were placed back into the coding queue for recoding. That is, they were rescored by coders who, under the standard-comparison method, had been deemed reliable against the weekly standard set of reliability cases.

Because two approaches were used for ongoing reliability assessment during the kindergarten 2006 coding operation, PreLAS and fine motor scores were determined to be reliable (and, therefore, included in the data file) through several means. All cases scored by the lead coder as adjudicator, and by the lead coder and principal investigator through consensus coding, were considered to be reliable. Cases in the inter-rater reliability sample for which two coders' scores agreed were considered to be reliable. Cases from the inter-rater reliability sample that were adjudicated by the lead coder also were considered to be reliable. Cases scored by a coder during a period in which the coder was determined to be reliable (based on the inter-rater reliability model) were included as reliable. Finally, cases completed by coders who had maintained reliability using the standard-comparison approach also were considered to be reliable. Table 73 summarizes the sources of scores for the Let's Tell Stories data in the kindergarten 2006 wave. The sources of scores for the fine motor data are summarized in table 74. Section 8.4 includes a discussion of how these changes in reliability methods may affect scores both within and across waves.

Table 73. Sources of scores for the Let's Tell Stories items, ECLS-B kindergarten 2006 data collection: 2006–07

Source	Number of cases coded	Percent of coded cases
All cases	6,850	100.0
Inter-rater reliability approach		
Cases adjudicated by lead coder	50	0.9
Cases checked by inter-rater reliability approach	50	0.9
Standard-comparison approach		
Cases completed through consensus coding	100	1.8
Cases checked by weekly reliability assessment	6,600	96.4

NOTE: The number of cases has been rounded to the nearest 50. Details may not sum to the total due to rounding. The total number of cases completed through consensus coding reflects the number of cases actually coded through consensus, regardless of whether they were used as part of the weekly reliability check or not. For example, two sets of reliability cases generally were prepared each week, even though many weeks only a single set was needed to check reliability.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 data collection, 2006–07.

Table 74. Sources of scores for the fine motor items, ECLS-B kindergarten 2006 data collection: 2006–07

	Square		Triangle		Asterisk		Circle-square	
Source	*n*	Percent	*n*	Percent	*n*	Percent	*n*	Percent
All cases	6,900	100.0	6,900	100.0	6,900	100.0	6,900	100.0
Inter-rater reliability approach								
Cases adjudicated by lead coder	100	1.5	100	1.3	50	1.0	100	1.7
Cases checked by inter-rater reliability approach	3,750	54.4	3,900	56.4	3,900	56.7	3,550	51.4
Standard-comparison approach								
Cases completed through consensus coding	300	4.2	300	4.2	300	4.2	300	4.2
Cases checked by weekly reliability assessment	2,800	39.9	2,650	38.1	2,650	38.1	2,950	42.7

NOTE: The number of cases (*n*) for each figure is rounded to the nearest 50. Details may not sum to the total due to rounding. The total number of cases completed through consensus coding reflects the number of cases actually coded through consensus, regardless of whether they were used as part of the weekly reliability check or not. For example, two sets of reliability cases generally were prepared each week, even though many weeks only a single set was needed to check reliability.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 data collection, 2006–07.

8.3 Coding Operations During the Kindergarten 2007 Data Wave

The procedures for coder training, certification, and ongoing assessment of reliability in the kindergarten 2007 wave were modeled after the standard-comparison approach adopted midway through the course of the kindergarten 2006 coding operations (see section 8.2.4).

8.3.1 Hiring and Training Coders

Three coders, one of whom was a returning coder from the kindergarten 2006 wave, were selected through an outside employment service to conduct coding for the kindergarten 2007 wave. They attended training November 25–29, 2007, which incorporated instruction and certification on both the PreLAS stories and the fine motor drawings. As in the preschool and kindergarten 2006 waves, all of the coders had a bachelor's degree in psychology, child development, early childhood education, sociology, or a related discipline. Preference was given to applicants with experience in data coding or child assessment. Training was planned and led by the RTI principal investigator with the assistance of the coding task leader and the lead coder. Before training, the scoring manual used for the kindergarten 2006 wave was fine-tuned, where needed, based on experiences in the kindergarten 2006 collection.

The RTI principal investigator and lead coder, who jointly oversaw both initial and ongoing coder reliability checks during the kindergarten 2006 wave, continued these responsibilities for the kindergarten 2007 wave. Prior to training, the reliability of both the RTI principal investigator and lead coder was verified (i.e., their continued agreement with the published scoring guidelines and lack of drift in the interim was confirmed). A sample of 40 audiotapes and Child Assessment Booklets, which were previously coded and assigned consensus scores by the lead coder and principal investigator during the kindergarten 2006 data collection, were selected for this purpose. Because these items and stories had been used during weekly coder reliability checks, each had a written justification for assigned scores and transcripts for the PreLAS stories on which the scores were based. After review of several practice cases, both the RTI principal investigator and lead coder separately scored a set of 10 cases selected by the coding task leader; this scoring was done "blind" (i.e., without the benefit of the earlier scores and accompanying feedback materials). Scores were sent back to the coding task leader for comparison with the standard scores previously assigned. The principal investigator and the lead coder each had to achieve at least a 90 percent agreement with the standard scores on each PreLAS story and each fine motor item. A second opportunity for recertification, using a new sample of 10 cases, was provided for any item for which 90 percent agreement was not achieved. The principal investigator achieved 100 percent agreement on all items; the lead coder required a second attempt for only one of the stories. The principal investigator and lead coder both achieved 100 percent agreement on all fine motor items.

During training of the three coders hired to code the kindergarten 2007 PreLAS and fine motor items, the principal investigator provided an overview of the items to be scored. Exemplars from the previously coded kindergarten 2006 standard cases were provided for each

story and each fine motor item. Coders then had time during training to practice the coding scales (individually and in groups) using cases drawn from these exemplars.

8.3.2 Certification of Field Coders

To ensure coding reliability for the Let's Tell Stories and fine motor items, the same item-level criterion for initial coder reliability utilized at kindergarten 2006 was applied at kindergarten 2007. Coders were certified on PreLAS scoring using a set of 10 standard cases, each consisting of 2 PreLAS stories (20 items total), and certified on each of 5 fine motor items using a set of 10 standard cases for each (50 items total). A set of 65 audiotapes and Child Assessment Booklets from recently completed kindergarten 2007 child assessments were used for certification. As in preschool and kindergarten 2006, coders were trained to score the stories using a 5-point scale, in addition to a "0" score applied for no response and for non-English responses to English stimuli. Fine motor items again were scored as pass or fail. As in the kindergarten 2006 wave, coders had to achieve 90 percent agreement per item against the standard scores to be certified. When the coder demonstrated 90 percent agreement or higher on both stories, the coder was considered certified and ready to code Let's Tell Stories. Coders also had to demonstrate 90 percent agreement or higher on each of the fine motor items.

For those coders who did not meet the criteria for certification on one or more items on the first attempt, additional consensus building instruction and practice were provided. These coders were then given a second set of certification items, drawn from additional Child Assessment Booklets and audiotapes set aside for this purpose, for those stories or fine motor items for which they failed to reach the 90 percent criterion on their first attempt. For example, if a coder was certified on story 1 but not story 2, that coder received additional instruction on story 2 and then scored 10 new story 2 cases. Percent agreement was again calculated. Three opportunities were provided for certification. The veteran coder from the kindergarten 2006 wave achieved a 100 percent agreement level on both Let's Tell Stories items on the first attempt. The two coders new to the task each passed one story on the first attempt. One passed the other story on the second attempt. The third coder passed the other story on the third attempt. One coder passed all fine motor items with a score of 90 percent or higher on the first attempt. The remaining two coders required a second attempt on the square only, having passed the other four fine motor items on their first attempt. For personal reasons, however, one coder withdrew from the job immediately after certification and never actually coded any live cases.

8.3.3 Reliability of Coding

Ongoing assessment of reliability in the kindergarten 2007 wave involved the same standard-comparison model adopted during the course of the kindergarten 2006 coding operations. With this approach, each week coders completed a common set of cases (i.e.,

reliability cases) for each item for which standard scores[125] had been developed by the lead coder and RTI principal investigator. These cases were randomly selected from among incoming cases not yet scored by coders. Coders were considered to be reliable if their percent agreement with the standard scores was 90 percent or greater on the set of reliability cases for each item.[126] Coders completed this set of cases on their first day of coding each week and received immediate feedback on their performance. If a coder achieved 90 percent or greater agreement on the set of common cases, that coder continued scoring. If a coder did not achieve 90 percent agreement or greater, that coder was not allowed to code until he or she had passed a second set of reliability cases for the failed item(s). Thus, if a coder did not meet criteria for Let's Tell Stories items, the coder would complete a second set of 10 Let's Tell Stories items (5 cases, 2 stories each) and if the coder did not meet criteria for one or more fine motor items the coder would complete a second set of 10 cases for those fine motor items.

As in the kindergarten 2006 wave, all coding staff participated in a weekly staff meeting at which standard scores for the first set of reliability cases were discussed, the rationale for each standard score was provided, and additional scoring guidance was given, if necessary. Following this meeting, coders who had met the 90 percent agreement criterion returned to production coding, while coders who had not reached that criterion were given the opportunity to receive more feedback and attempt a second set of reliability cases. The performance of coders on the weekly reliability cases was generally quite strong for the PreLAS items. Percent agreement across all coders is shown in table 75.

Table 75. Coder percent agreement with standard scores for Let's Tell Stories items using a standard-comparison model, ECLS-B kindergarten 2007 data collection: 2007–08

Item	Number of coders	Number of reliability cases	Number of cases reviewed using this method	Average percent agreement	Percent of coders achieving agreement between 90 and 100 percent	Range of percent agreement
PreLAS stories	2	75	1,741	92.3	100	90–100

NOTE: Range of percent agreement includes reliability sets through which coders were certified.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2007 data collection, 2007–08.

The performance of coders on the weekly reliability cases also was generally quite strong for the fine motor items. Percent agreement across all coders is shown for each fine motor item in table 76.

[125] The standard scores were determined by consensus coding; the lead coder and principal investigator independently scored the tapes and adjudicated differences through discussion. In this way, a set of reliability tapes was created.
[126] To measure reliability each week, coders completed 10 cases for each fine motor item, and 5 Let's Tell Stories cases with 2 stories each (or 10 items in total). This resulted in percent agreement values for each of the fine motor items, and a single value for Let's Tell Stories (i.e., there were not separate estimates for each story).

Table 76. Coder percent agreement with standard scores for fine motor item coding using a standard-comparison model, ECLS-B kindergarten 2007 data collection: 2007–08

Item	Number of coders	Number of reliability cases	Number of cases reviewed using this method	Average percent agreement	Percent of coders achieving agreement between 90 and 100 percent	Range of percent agreement
Square	2	210	1,687	96.9	100	90–100
Triangle	2	210	1,687	95.0	100	90–100
Asterisk	2	210	1,687	97.3	100	90–100
Circle-square	2	210	1,687	95.0	100	90–100

NOTE: Range of percent agreement includes reliability sets through which coders were certified.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2007 data collection, 2007–08.

8.3.4 Source of Scores for the Kindergarten 2007 Wave

All cases scored by the lead coder and RTI principal investigator during reliability coding were considered to be reliable and included in the data set. Cases completed by coders who had maintained reliability using the standard-comparison approach also were considered to be reliable and were included. Table 77 summarizes the sources of scores for the Let's Tell Stories data in the kindergarten 2007 wave. The sources of scores for the fine motor items are provided in table 78.

Table 77. Sources of scores for the Let's Tell Stories items, ECLS-B kindergarten 2007 data collection: 2007–08

Source	Number of cases coded	Percent of coded cases
All cases	1,900	100.0
Cases completed through consensus coding	150	9.0
Cases checked by weekly reliability assessment	1,750	91.0

NOTE: The number of cases has been rounded to the nearest 50. The details may not sum to the total due to rounding. The total number of cases completed through consensus coding reflects the number of cases actually coded through consensus, regardless of whether they were used as part of the weekly reliability check or not. For example, two sets of reliability cases generally were prepared each week, even though many weeks only a single set was needed to check reliability.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2007 data collection, 2007–08.

Table 78. Sources of scores for the fine motor items, ECLS-B kindergarten 2007 data collection: 2007–08

	Square		Triangle		Asterisk		Circle-square	
Source	*n*	Percent	*n*	Percent	*n*	Percent	*n*	Percent
All cases	1,950	100.0	1,950	100.0	1,950	100.0	1,950	10.00
Cases completed through consensus coding	250	12.9	250	12.9	250	12.9	250	12.9
Cases checked by weekly reliability assessment	1,700	87.1	1,700	87.1	1,700	87.1	1,700	87.1

NOTE: The number of cases (*n*) for each figure has been rounded to the nearest 50. The details may not sum to the total due to rounding. The total number of cases completed through consensus coding reflects the number of cases actually coded through consensus, regardless of whether they were used as part of the weekly reliability check or not. For example, two sets of reliability cases generally were prepared each week, even though many weeks only a single set was needed to check reliability.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2007 data collection, 2007–08.

8.4 Impact of Differences in Coding Operations Across Waves

As described in this chapter, the methods used to assess the reliability of coding across data waves, and within the kindergarten 2006 wave, were changed over time. The changes in methodology were driven by the need to maintain the highest levels of data quality given operational considerations for completing data collection and coding operations. The most dramatic change in approaches occurred between the preschool and kindergarten 2006 data collections. Between these two waves, the criteria for agreement were made more stringent, with a more precise match required for all items (exact match rather than "within 1 point" for Let's Tell Stories during preschool), and the percentage agreement criteria also increased. Importantly, across waves of data collection in the ECLS-B, the standards for reliability approximated or exceeded generally accepted standards in the field for reliability estimates.

The result of the increases in these criteria is that the error associated with each score would be expected to be smaller in the kindergarten waves than in the preschool wave because reliability was more stringently monitored. The differences in the reliability criteria in certification as well as ongoing reliability assessment across waves was not proportionally large (an increase as large as 10 points in the expectation and in realization) but are real. Thus, the standard errors associated with kindergarten scores are generally lower than those of the preschool scores. This is problematic when considering scores on the same items administered across waves, where analysts might be interested in examining change in scores over time to show growth. Using individual items in this way is inherently risky because individual item scores are generally not as stable as aggregate scores (i.e., overall copy forms or Let's Tell Stories score) over time. When the reliability criteria are also changing over time, it is likely that an individual child's scores on any given item will show some variability, possibly including an apparent decrease in performance, while aggregate scores (either across children, or across similar items completed by the same child) will be more stable over time. Therefore,

consideration should be given to the use of aggregate, rather than individual story or fine motor scores in analysis.

References

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

Andreassen, C., Fletcher, P., and West, J. (2005). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Methodology Report for the 9-Month Data Collection (2001–02), Volume 1: Psychometric Characteristics* (NCES 2005-100). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Brooks-Gunn, J., Liaw, F.R., Michael, R.T., and Zamsky, E.S. (1992). *Manual for Coding Freeplay Parenting Styles: From the Newark Observational Study of the Teenage Parent Demonstration*. Unpublished coding scales. New York: Teachers College, Columbia University.

Bruininks, R.H. (1978). *Bruininks-Oseretski Test of Motor Proficiency*. Circle Pines, MN: American Guidance Service.

Brush, L., Salinger, T., Sussman, A., and Kirshstein, R. (2003). *Cognitive Assessment Plan for the ECLS-B Preschool Battery.* Prepared for the National Center for Education Statistics, U.S. Department of Education. Washington, DC: American Institutes for Research.

Burns, L.J., Heuer, R., Ingels, S.J., Pollack, J., Pratt, D.J., Rock, D., Rogers, J., Scott, L.A., Siegel, P., and Stutts, E. (2003). *Education Longitudinal Study: 2002 Field Test Report* (NCES 2003-03). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved December 5, 2006, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200303.

Cole, N.S., and Moss, P.A. (1989). Bias in Test Use. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 201–219). New York: American Council on Education/Macmillan.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood From Incomplete Data Via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society*, *39*, Series B, 1–38.

Dorans, N.J., and Kulick, E. (2006). Differential Item Functioning on the MMSE: An Application of the Mantel-Haenzel and Standardization Procedures. *Medical Care, 44 S3*, S107-S114.

Duncan, S.E., and De Avila, E.A. (1998). *PreLAS 2000.* Monterey, CA: CTB/McGraw-Hill.

Dunn, L.M., and Dunn, L.M. (1997). *Peabody Picture Vocabulary Test–Third Edition* (PPVT-III). Upper Saddle River, NJ: Pearson Publishing. Accessed October 30, 2006, from http://ags.pearsonassessments.com/group.asp?nGroupInfoID=a12010.

Dunn, L., Padilla, E., Lugo, D., and Dunn, L. (1986). *TVIP: Test de Vocabulario en Imagenes Peabody*. Upper Saddle River, NJ: Pearson Publishing.

Fauth, R.C., Brady-Smith, C., and Brooks-Gunn, J. (2003). *Parent-Child Interaction Rating Scales for the Play Doh Task and Father-Child Interaction Rating Scales for the Three-Bag Task*. New York: National Center for Children and Families (NCCF), Teachers College, Columbia University.

Fenson, L., Dale, P.S., Reznick, J.S., Bates, E., Thal, D.J., and Pethick, S.J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development, 59*(5, Serial No. 242).

Freeman, D. (1968). *Corduroy.* New York: Penguin Putnam Books for Young Readers.

Ginsburg, H.P., and Baroody, A.J. (2003). *Test of Early Mathematics Ability* (3rd ed.). Austin, TX: PRO-ED, Inc.

Goodson, B.D., Layzer, C.I., Smith, W.C., and Rimzdius, T. (2004). *Observation Measure of Language and Literacy Instruction (OMLIT)*. Cambridge, MA: Abt Associates, Inc

Gresham, F.M., and Elliott, S.N. (1990). *Social Skills Rating System Manual*. Circle Pines, MN: American Guidance Service.

Henderson, S.E., and Sugden, D.A. (1992). *Movement Assessment Battery for Children*. London: Psychological Corporation.

Holland, P.W., and Thayer, D.T. (1986). *Differential Item Function and the Mantel-Haenszel Procedure* (ETS Research Report No. 86-31). Princeton, NJ: Educational Testing Service.

Ketchie, B., Lang, N., Brush, L., and Kirstein, R. (2003). *Recommended Physical Assessment Instrument for the ECLS-B Preschool Battery: Results From the Spring Pilot Test.* Washington, DC: American Institutes for Research.

Kirsch, I.S., Jungeblut, A., Jenkins, L., and Kolstad, A. (1993). *Adult Literacy in America: A First Look at the Findings of the National Adult Literacy Survey* (NCES 93-275). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved December 5, 2006, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=93275.

Leventhal, F. (1998). *The Need for Early Intervention: Development of a Language Skills Behavioral Checklist for Preschool Children*. Thesis (Ph.D.), Columbia University.

Lonigan, C.J., Wagner, R.K., Torgesen, J.K., and Rashotte, C.A. (2002). *Preschool Comprehensive Test of Phonological & Print Processing.* Unpublished assessment.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Publishers.

Mantel, N., and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *Journal of the National Cancer Institute, 22*: 719–748.

Meisels, S.J., Marsden, D.B., Wiske, M.S., and Henderson, L.W. (1997). *The Early Screening Inventory–Revised (ESI-R)*. New York: Pearson Early Learning.

Merrell, K.M. (2003). *Preschool and Kindergarten Behavior Scales, Second Edition (PKBS-2)*. Retrieved October 27, 2006, from http://www.proedinc.com/Scripts/prodView.asp?idProduct=2285.

Mislevy, R.J., and Bock, R.D. (1982). *BILOG: Item Analysis and Test Scoring With Binary Logistic Models* [Computer software]. Mooresville, IN: Scientific Software.

Mislevy, R.J., Johnson, E.G., and Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics*, *17*(2): 131–154.

Muraki, E.J., and Bock, R.D. (1987). *BIMAIN: A Program for Item Pool Maintenance in the Presence of Item Parameter Drift and Item Bias* [Computer software]. Mooresville, IN: Scientific Software.

Muraki, E.J., and Bock, R.D. (1991). *PARSCALE: Parameter Scaling of Rating Data* [Computer software]. Chicago, IL: Scientific Software, Inc.

Nathanson, L., Lang, N., Than, V., Ketchie, B., Brush, L., and Kirshstein, R. (2003). *Recommended Cognitive Assessment Instrument for the ECLS-B Preschool Battery: Results of the 2003 Pilot Test.* Prepared for the National Center for Education Statistics, U.S. Department of Education. American Institutes for Research, Washington, DC.

Owings, J. (1995). *Psychometric Report for the NELS:88 Base Year Through Second Follow-up* (NCES 95-382). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved December 5, 2006, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=95382.

Pollack, J., Atkins-Burnett, S., Najarian, M., and Rock, D. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for the Fifth Grade* (NCES 2006-036rev). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved December 5, 2006, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006036rev.

Rathmann, P. (1994). *Good Night, Gorilla.* New York: G. P. Putnam's Sons.

Rock, D.A., Hilton, T.L., Pollack, J.M., Ekstrom, R.B., and Goertz, M.E. (1985). *Psychometric Analysis of the NLS-72 and the High School and Beyond Test Batteries* (NCES 85-217). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Rock, D.A., and Pollack, J. (1987). The Cognitive Test Battery. In S.J. Ingels, et al., *Field Test Report: National Education Longitudinal Study of 1988 (Base Year).* Chicago: NORC, University of Chicago.

Rock, D.A., and Pollack, J. (2002). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for Kindergarten Through the First Grade* (NCES 2002-05). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved December 5, 2006, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200205.

Snow, K., Derecho, A., Wheeless, S., Lennon, J., Rosen, J., Rogers, J., Kinsey, S., Morgan, K., and Einaudi, P. (2009). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Kindergarten 2006 and 2007 Data File User's Manual* (NCES 2010-010). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Snow, K., Thalji, L., Derecho, A., Wheeless, S., Lennon, J., Kinsey, S., Rogers, J., Raspa, M., and Park, J. (2007). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Preschool Data File User's Manual (2005–06)* (NCES 2007-055). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Wallace, I., and Dedek, J. (2005). *Early Childhood Longitudinal Study, Birth Cohort: The Preschool Year (ECLS-B, Preschool Year) Field Test Report #5, Child Assessment*. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Ware, A., Brady, C., O'Brien, C., and Berlin, L.J. (1998). *14-Month Child-Parent Interaction Rating Scales for the Three Bag Assessment*. New York: Center for Children and Families, Teachers College, Columbia University.

West, J., Denton, K., and Germino-Hausken, E. (2000). *America's Kindergartners* (NCES 2000-070). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Yamamoto, K., and Mazzeo, J. (1992). Item Response Theory: Scale Linking in NAEP. *Journal of Education Statistics*, *17*: 155–173.

Appendix A: Abbreviations

ACASI	audio computer-assisted self-interviewing
AIR	American Institutes for Research
BMI	body mass index
BSF-R	Bayley Short Form—Research Edition
CAPI	computer-assisted personal interviewing
CCD	Common Core of Data
CCO	Child Care Observation
CLIO	Classroom Literacy Interventions and Outcomes Study
DIF	differential item functioning
EAP	expected a posteriori
ECB	electronic codebook
ECEP	early care and education provider
ECLS-B	Early Childhood Longitudinal Study, Birth Cohort
ECLS-K	Early Childhood Longitudinal Study, Kindergarten Class of 1998–99
EHS	Early Head Start
EM	expectation maximization
ESI-K	Early Screening Inventory-Kindergarten
ESI-R	Early Screening Inventory–Revised
ETS	Educational Testing Service
FACES	Family and Child Experiences Survey
FI	field interviewer
IES	Institute of Education Sciences
IRT	item response theory
M-CDI	MacArthur Communicative Development Inventory
M-H	Mantel-Haenszel
MUAC	middle upper arm circumference
NAEP	National Assessment of Educational Progress
NCES	National Center for Education Statistics
NHANES	National Health and Nutrition Examination Survey
OMLIT	Observation Measure of Language and Literacy Instruction
P-DIF	primary item discrepancy index

PI	principal investigator
PKBS-2	Preschool and Kindergarten Behavior Scales
PPVT-III	Peabody Picture Vocabulary Test–Third Edition
Pre-CTOPPP	Preschool Comprehensive Test of Phonological and Print Processing
PSAQ	Parent Self-Administered Questionnaire
PSS	Private School Universe Survey
RAP	Read Aloud Profile
RAPT	Reading Aloud Profile–Together
RFSAQ	Resident Father Self-Administered Questionnaire
SD	standard deviation
SSRS	Social Skills Rating System
TEMA-3	Test of Early Mathematical Ability-3
TRP	Technical Review Panel
TSAQ	Teacher Self-Administered Questionnaire
TVIP	Test de Vocabulario en Imagenes Peabody
WECEP	wrap-around early care and education provider

Appendix B: ECLS-B Item Parameters and Item Fit by Waves

This appendix provides the IRT parameters, and observed and predicted probability correct (P+) values, for items on the ECLS-B early reading (table B-1) and mathematics items (table B-2).

Table B-1. Early reading item parameters and fit by wave

Item	Test form PK	Test form K06 & K07	IRT parameters			PK				K06				K07			
							P+				P+				P+		
			a	b	c	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff
POURIN P	Lang	†	0.51	-1.51	0.15	8208	0.75	0.74	0.01	†	†	†	†	†	†	†	†
Empty	Lang	†	0.76	-1.67	0.07	8260	0.80	0.80	0.00	†	†	†	†	†	†	†	†
Peeking	Lang	†	0.70	-1.03	0.06	8227	0.66	0.66	0.00	†	†	†	†	†	†	†	†
Catpil	Lang	†	0.68	-1.52	0.10	8080	0.77	0.77	0.00	†	†	†	†	†	†	†	†
Jogging	Lang	†	0.56	0.13	0.08	7731	0.44	0.43	0.01	†	†	†	†	†	†	†	†
Statue	Lang	†	0.69	-0.28	0.06	7489	0.51	0.50	0.01	†	†	†	†	†	†	†	†
Liquid	Lang	†	0.43	0.50	0.11	7053	0.43	0.43	0.01	†	†	†	†	†	†	†	†
Root	Lang	†	0.75	1.36	0.26	6565	0.37	0.36	0.01	†	†	†	†	†	†	†	†
Pair	Lang	†	0.71	1.55	0.27	5969	0.37	0.36	0.01	†	†	†	†	†	†	†	†
Directin	Lang	†	0.65	2.09	0.28	5475	0.34	0.34	0.01	†	†	†	†	†	†	†	†
R	Lit	†	2.14	-0.39	0.41	8119	0.70	0.69	0.01	†	†	†	†	†	†	†	†
S	Lit	†	1.78	-0.40	0.28	8168	0.63	0.62	0.01	†	†	†	†	†	†	†	†
E	Lit	†	1.79	-0.43	0.24	8156	0.62	0.61	0.01	†	†	†	†	†	†	†	†
H	Lit	†	1.76	-0.27	0.00	7650	0.46	0.43	0.03	†	†	†	†	†	†	†	†
Tuh	Lit	†	1.61	-0.06	0.12	3191	0.68	0.67	0.01	†	†	†	†	†	†	†	†
Kuh	Lit	†	1.48	-0.07	0.00	2456	0.71	0.68	0.03	†	†	†	†	†	†	†	†
He	Lit	†	1.13	1.64	0.00	1912	0.13	0.12	0.02	†	†	†	†	†	†	†	†
Mop	Lit	†	2.46	0.89	0.43	7749	0.47	0.46	0.01	†	†	†	†	†	†	†	†
BEG BI P	Lit	†	1.48	0.05	0.22	7790	0.47	0.46	0.01	†	†	†	†	†	†	†	†
Shoe P	Lit	†	0.46	-1.78	0.13	7612	0.80	0.76	0.04	†	†	†	†	†	†	†	†
Battle	Lit	†	0.41	0.00	0.17	7303	0.56	0.52	0.03	†	†	†	†	†	†	†	†
PlaygrnP	Lit	†	1.05	0.32	0.00	7445	0.29	0.26	0.03	†	†	†	†	†	†	†	†
Was	Lit	†	1.05	0.13	0.41	8063	0.60	0.59	0.01	†	†	†	†	†	†	†	†
TedTitle	Lit	†	0.81	-0.53	0.00	7843	0.55	0.52	0.03	†	†	†	†	†	†	†	†
StrtRead	Lit	†	0.83	-0.13	0.00	8025	0.42	0.40	0.02	†	†	†	†	†	†	†	†
GoNext	Lit	†	0.85	0.46	0.00	7910	0.27	0.25	0.01	†	†	†	†	†	†	†	†
TedGo	Lit	†	0.87	0.73	0.00	7912	0.20	0.19	0.01	†	†	†	†	†	†	†	†
BEG R 6	†	R	2.50	0.34	0.12	†	†	†	†	6696	0.59	0.59	0.00	1850	0.82	0.80	0.02
BEG B 6	†	R	1.22	0.74	0.05	†	†	†	†	6726	0.43	0.42	0.01	1856	0.57	0.59	-0.02
END L 6	†	R	1.77	0.65	0.13	†	†	†	†	3909	0.67	0.67	0.00	1494	0.80	0.76	0.04
END F 6	†	R	1.19	0.70	0.08	†	†	†	†	3983	0.62	0.61	0.00	1514	0.70	0.69	0.01

See notes at end of table.

Table B-1. Early reading item parameters and fit by wave—Continued

Item	Test form PK	Test form K06 & K07	IRT parameters			PK				K06				K07			
							P+				P+				P+		
			a	b	c	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff
END P 6	†	R	1.57	0.86	0.08	†	†	†	†	4003	0.56	0.56	0.01	1536	0.64	0.65	0.00
END D 6	†	R	1.50	1.06	0.07	†	†	†	†	3966	0.47	0.47	0.00	1519	0.57	0.56	0.02
WENT 6	†	R	2.37	1.48	0.00	†	†	†	†	3467	0.26	0.24	0.02	1377	0.34	0.32	0.03
DOWN 6	†	R	2.67	1.67	0.00	†	†	†	†	3471	0.18	0.16	0.02	1370	0.24	0.23	0.01
JEEP 6	†	R	2.34	1.63	0.00	†	†	†	†	3504	0.20	0.19	0.02	1410	0.26	0.25	0.01
BACKPA 6	†	R	3.07	2.02	0.23	†	†	†	†	634	0.49	0.48	0.01	372	0.55	0.50	0.05
LISTEN 6	†	R	4.91	2.10	0.26	†	†	†	†	611	0.47	0.45	0.02	348	0.49	0.47	0.02
RIDEBI 6	†	R	3.73	2.13	0.22	†	†	†	†	620	0.44	0.42	0.03	375	0.40	0.43	-0.03
SIZES 6	†	R	3.97	2.15	0.20	†	†	†	†	577	0.40	0.40	0.00	360	0.41	0.41	0.00
STORYE 6	†	0, 1	1.56	0.44	0.11	†	†	†	†	4479	0.40	0.38	0.01	820	0.55	0.52	0.03
QMARK 6	†	1	1.01	0.97	0.03	†	†	†	†	2229	0.33	0.31	0.02	588	0.41	0.35	0.06
KAYLAF 6	†	1, 2	0.39	0.33	0.00	†	†	†	†	4345	0.60	0.58	0.02	1608	0.68	0.62	0.06
COULDN 6	†	1, 2	0.39	0.90	0.00	†	†	†	†	3953	0.56	0.50	0.06	1495	0.63	0.53	0.10
COULD 6	†	1, 2	0.28	1.63	0.00	†	†	†	†	3972	0.47	0.41	0.06	1497	0.49	0.44	0.05
AWARDI 6	†	1, 2	0.40	0.67	0.15	†	†	†	†	4427	0.60	0.60	0.00	1629	0.68	0.63	0.05
JOGGIN 6	†	1, 2	0.63	0.56	0.10	†	†	†	†	4570	0.60	0.61	-0.01	1664	0.69	0.66	0.03
YELLOW 6	†	1, 2	1.44	0.86	0.00	†	†	†	†	4152	0.51	0.50	0.02	1555	0.68	0.60	0.09
YOU 6	†	2	2.64	1.17	0.00	†	†	†	†	2014	0.60	0.60	0.01	971	0.71	0.65	0.06
FISHIN 6	†	2	3.47	1.77	0.00	†	†	†	†	1738	0.26	0.23	0.03	855	0.28	0.27	0.00
THROUG 6	†	2	3.71	2.21	0.00	†	†	†	†	1655	0.10	0.08	0.01	797	0.10	0.11	-0.01
CATCH 6	†	2	2.88	1.69	0.00	†	†	†	†	1769	0.32	0.29	0.03	870	0.33	0.33	0.00
TOIL 6	†	2	1.87	2.36	0.00	†	†	†	†	1722	0.12	0.10	0.02	861	0.10	0.12	-0.02
mike	†	2	0.70	1.46	0.00	†	†	†	†	2041	0.48	0.46	0.02	982	0.52	0.49	0.04
tiger	†	2	1.12	2.23	0.00	†	†	†	†	2032	0.21	0.19	0.02	969	0.20	0.21	-0.01
winter	†	2	1.02	2.30	0.00	†	†	†	†	2061	0.20	0.18	0.01	983	0.21	0.21	0.00
ORSAT 6	†	2	1.56	1.38	0.00	†	†	†	†	2027	0.49	0.47	0.02	974	0.54	0.51	0.03
ORPIG 6	†	2	1.24	1.25	0.00	†	†	†	†	2053	0.55	0.53	0.02	983	0.59	0.57	0.02
ORTAIL 6	†	2	1.61	1.45	0.00	†	†	†	†	2041	0.45	0.43	0.02	976	0.49	0.48	0.02
ORHAND 6	†	2	1.71	1.56	0.00	†	†	†	†	2003	0.39	0.37	0.02	968	0.44	0.41	0.03
BOYBIR 6	†	2	3.48	1.67	0.20	†	†	†	†	1611	0.47	0.45	0.02	814	0.47	0.48	-0.01
CANINB 6	†	2	1.99	1.82	0.26	†	†	†	†	1597	0.48	0.46	0.02	832	0.53	0.48	0.05
KITNBE 6	†	2	3.02	1.70	0.20	†	†	†	†	1504	0.45	0.45	0.01	797	0.50	0.47	0.04

See notes at end of table.

Table B-1. Early reading item parameters and fit by wave—Continued

Item	Test form PK	Test form K06 & K07	IRT parameters			PK				K06				K07			
							P+				P+				P+		
			a	b	c	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff
GIRLRE 6	†	2	2.55	1.92	0.26	†	†	†	†	1581	0.44	0.41	0.03	850	0.44	0.43	0.01
KIMCAT 6	†	2	1.58	1.81	0.38	†	†	†	†	532	0.72	0.70	0.01	353	0.73	0.70	0.03
NEEDHO 6	†	2	2.93	1.61	0.16	†	†	†	†	516	0.74	0.73	0.01	346	0.72	0.71	0.01
LIKEDR 6	†	2	2.24	2.22	0.19	†	†	†	†	499	0.44	0.42	0.01	338	0.41	0.42	-0.01
LETRF P6	Lit	R	2.15	-0.30	0.00	7049	0.47	0.47	0.00	6684	0.79	0.76	0.03	1857	0.92	0.92	0.00
LETRD P6	Lit	R	1.89	-0.34	0.00	5359	0.59	0.60	0.00	6675	0.79	0.77	0.02	1850	0.93	0.92	0.01
LETRM P6	Lit	R	1.81	-0.35	0.00	4215	0.68	0.69	-0.01	6689	0.78	0.76	0.02	1860	0.93	0.91	0.02
LETRT P6	Lit	R	1.89	-0.20	0.00	3603	0.63	0.68	-0.04	6648	0.76	0.72	0.04	1851	0.92	0.90	0.03
Nn P6	Lit	0, 1	1.51	-0.04	0.22	3186	0.70	0.70	0.00	4514	0.64	0.62	0.01	836	0.76	0.75	0.00
S2 P6	Lit	0	1.03	-0.47	0.00	2866	0.74	0.75	-0.01	2106	0.52	0.46	0.06	194	0.70	0.53	0.17
BEG P P6	Lit	R	1.98	0.45	0.10	2113	0.49	0.50	-0.02	6714	0.55	0.54	0.01	1866	0.75	0.74	0.01
BEG L P6	Lit	R	2.51	0.46	0.13	2085	0.44	0.50	-0.07	6685	0.58	0.55	0.02	1860	0.78	0.76	0.03
RUNS P6	Lit	R	2.01	1.39	0.00	619	0.28	0.18	0.10	3472	0.30	0.28	0.02	1342	0.40	0.37	0.02
truck P6	Lit	1	0.92	0.29	0.43	7348	0.62	0.59	0.03	2355	0.75	0.74	0.01	622	0.82	0.77	0.05
sneezeP6	Lit	1	1.62	0.25	0.22	7614	0.40	0.40	-0.01	2396	0.71	0.69	0.03	629	0.79	0.74	0.05
sock P6	Lit	1	1.59	0.34	0.39	7608	0.52	0.52	0.01	2409	0.74	0.72	0.02	641	0.83	0.76	0.07
Run P6	Lit	1	1.61	0.50	0.35	7724	0.46	0.45	0.00	2357	0.67	0.65	0.03	632	0.76	0.69	0.07
AirporP6	Lit	1	1.14	0.58	0.00	7425	0.20	0.19	0.01	2354	0.45	0.43	0.03	633	0.55	0.48	0.07
Heat P6	Lit	1	0.80	0.84	0.00	7239	0.20	0.19	0.01	2318	0.42	0.37	0.05	623	0.47	0.40	0.07
BEGIN P6	Lit	1, 2	1.05	0.31	0.14	5225	0.41	0.42	-0.01	4478	0.51	0.47	0.04	831	0.62	0.58	0.04
NEXTLIP6	Lit	1, 2	1.20	0.38	0.05	5162	0.32	0.33	-0.01	4457	0.42	0.38	0.03	814	0.57	0.51	0.06
BEGWORP6	Lit	1	1.38	0.95	0.07	5075	0.20	0.19	0.02	2309	0.32	0.30	0.02	617	0.33	0.34	-0.01

† Not applicable.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.

Table B-2. Mathematics item parameters and fit by wave

Item	Test Form PK	Test Form K06 & K07	IRT parameters			PK				K06				K07			
							P+				P+				P+		
			a	b	c	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff
LongLine	R	†	0.81	-1.99	0.29	8133	0.90	0.90	0.01	†	†	†	†	†	†	†	†
FewBallP	R	†	0.58	-0.82	0.00	8165	0.59	0.58	0.01	†	†	†	†	†	†	†	†
MoreAppl	R	†	0.34	-2.07	0.00	8157	0.72	0.71	0.01	†	†	†	†	†	†	†	†
Cookies	R	†	0.62	-0.57	0.00	8164	0.53	0.53	0.01	†	†	†	†	†	†	†	†
LastBikP	R	†	0.69	-0.64	0.18	8049	0.63	0.63	0.00	†	†	†	†	†	†	†	†
BtwnBike	R	†	0.73	-0.52	0.23	7856	0.64	0.63	0.01	†	†	†	†	†	†	†	†
Triang1	R	†	0.70	-0.95	0.17	7892	0.70	0.69	0.01	†	†	†	†	†	†	†	†
FourDucP	R	†	1.08	-0.18	0.32	7998	0.62	0.61	0.01	†	†	†	†	†	†	†	†
LBLSqA	R	†	0.92	-0.37	0.26	7928	0.63	0.62	0.01	†	†	†	†	†	†	†	†
Pattern	R	†	0.77	0.31	0.00	8108	0.32	0.31	0.01	†	†	†	†	†	†	†	†
Shapes3	R	†	0.95	0.56	0.00	8094	0.23	0.22	0.01	†	†	†	†	†	†	†	†
Stars4	R	†	1.50	-1.51	0.00	8168	0.84	0.84	0.00	†	†	†	†	†	†	†	†
3BANAN P	R	†	0.49	-0.95	0.18	7062	0.71	0.69	0.03	†	†	†	†	†	†	†	†
Fingers5	L	†	0.89	-1.46	0.00	2519	0.58	0.56	0.03	†	†	†	†	†	†	†	†
Books	L	†	1.96	-1.38	0.00	2446	0.60	0.57	0.03	†	†	†	†	†	†	†	†
Fish	L	†	1.87	-1.12	0.00	2495	0.44	0.41	0.02	†	†	†	†	†	†	†	†
Counter9	L	†	1.17	-1.45	0.00	2432	0.61	0.57	0.04	†	†	†	†	†	†	†	†
Cntrs5	L	†	1.32	-0.59	0.00	2396	0.24	0.21	0.03	†	†	†	†	†	†	†	†
NUMBE9 6	†	R	1.69	0.04	0.05	†	†	†	†	6731	0.66	0.67	-0.01	1863	0.84	0.82	0.02
STICK 6	†	R	1.12	0.13	0.27	†	†	†	†	6779	0.71	0.70	0.00	1865	0.81	0.81	-0.01
numbe6 6	†	R	1.64	0.68	0.02	†	†	†	†	6554	0.41	0.42	-0.01	1839	0.69	0.61	0.08
numb20 6	†	R	1.38	2.00	0.01	†	†	†	†	6403	0.09	0.09	0.00	1799	0.18	0.16	0.02
CARS5 6	†	R	1.37	0.99	0.08	†	†	†	†	6667	0.37	0.36	0.01	1848	0.50	0.52	-0.02
MARBLE 6	†	R	1.24	1.24	0.02	†	†	†	†	6531	0.26	0.26	0.00	1833	0.43	0.40	0.03
PENNIE 6	†	R	1.46	1.68	0.02	†	†	†	†	6642	0.15	0.14	0.01	1825	0.24	0.24	0.00
CHOCOL 6	†	0, 1, 2	1.16	-0.29	0.01	†	†	†	†	6598	0.75	0.74	0.01	1819	0.86	0.86	0.00
VANILL 6	†	0, 1, 2	1.18	-0.50	0.01	†	†	†	†	6626	0.80	0.79	0.01	1837	0.90	0.89	0.00
STRAWB 6	†	0, 1, 2	1.31	-0.67	0.12	†	†	†	†	1185	0.60	0.55	0.05	118	0.66	0.61	0.05
PBRUSH 6	†	0, 1, 2	1.23	0.04	0.25	†	†	†	†	6517	0.74	0.73	0.01	1832	0.86	0.84	0.02
SMLGSM 6	†	R	1.48	0.29	0.42	†	†	†	†	5453	0.80	0.80	0.00	1713	0.88	0.87	0.01
OOOX 6	†	1, 2	1.03	0.47	0.31	†	†	†	†	5282	0.72	0.70	0.01	1683	0.81	0.78	0.02
HALFOV 6	†	1, 2	0.75	0.74	0.28	†	†	†	†	5322	0.63	0.63	0.01	1677	0.73	0.70	0.03

See notes at end of table.

Table B-2. Mathematics item parameters and fit by wave—Continued

Item	Test Form PK	Test Form K06 & K07	IRT parameters			PK				K06				K07			
							P+				P+				P+		
			a	b	c	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff
Danny6	†	1	0.63	0.92	0.00	†	†	†	†	4049	0.39	0.38	0.02	1036	0.41	0.43	-0.02
Rachel6	†	1	1.88	0.82	0.25	†	†	†	†	3021	0.51	0.50	0.02	807	0.56	0.58	-0.02
CRAYS2 6	†	1, 2	1.47	1.38	0.13	†	†	†	†	4231	0.37	0.35	0.02	1429	0.45	0.47	-0.01
BUGS 6	†	1, 2	1.04	1.07	0.24	†	†	†	†	5251	0.52	0.52	0.00	1667	0.61	0.61	0.00
SHAPES 6	†	1, 2	0.54	1.50	0.26	†	†	†	†	1336	0.63	0.62	0.01	671	0.65	0.65	0.00
eclips 6	†	1, 2	0.92	1.34	0.20	†	†	†	†	1365	0.62	0.63	-0.01	679	0.70	0.68	0.03
7CIRCS 6	†	2	1.63	1.43	0.05	†	†	†	†	1226	0.56	0.55	0.01	629	0.62	0.63	-0.01
coun12 6	†	2	1.31	1.57	0.00	†	†	†	†	1325	0.44	0.44	-0.01	667	0.57	0.52	0.05
lunch 6	†	2	1.55	2.33	0.02	†	†	†	†	1300	0.15	0.15	0.00	655	0.23	0.20	0.03
Cars3 6	†	2	1.67	2.40	0.02	†	†	†	†	620	0.18	0.18	0.01	402	0.24	0.22	0.02
CANDI 6	†	2	1.92	2.40	0.03	†	†	†	†	635	0.17	0.16	0.01	407	0.20	0.20	0.00
number 6	†	2	1.35	2.10	0.01	†	†	†	†	1350	0.24	0.23	0.01	678	0.29	0.29	0.00
3plus4 6	†	2	2.61	1.47	0.04	†	†	†	†	1338	0.52	0.50	0.02	675	0.58	0.61	-0.02
7minu3 6	†	2	2.55	2.28	0.23	†	†	†	†	1118	0.34	0.31	0.04	606	0.32	0.34	-0.02
12plu6 6	†	2	2.13	2.06	0.02	†	†	†	†	1094	0.22	0.21	0.01	591	0.30	0.27	0.03
17min4 6	†	2	2.35	2.40	0.01	†	†	†	†	1077	0.11	0.09	0.02	588	0.12	0.13	-0.01
TallTrP6	R	0	0.55	-0.48	0.20	8153	0.61	0.61	0.00	1282	0.58	0.56	0.02	134	0.55	0.57	-0.03
MoreBaP6	R	0	0.69	-2.21	0.00	8201	0.86	0.85	0.00	1306	0.85	0.83	0.02	144	0.85	0.84	0.02
personP6	R	R	1.76	0.88	0.06	8065	0.13	0.15	-0.01	6776	0.38	0.36	0.02	1859	0.59	0.55	0.05
LG-SM-P6	R	R	1.05	-0.13	0.27	7859	0.59	0.56	0.03	6673	0.75	0.76	-0.01	1847	0.85	0.86	-0.01
4LINESP6	R	1, 2	0.84	0.73	0.20	7588	0.35	0.36	-0.02	5327	0.62	0.58	0.03	1705	0.72	0.67	0.05
Stars5P6	R	0, 1	1.41	-1.20	0.00	7705	0.77	0.78	-0.01	5442	0.91	0.90	0.01	1186	0.96	0.95	0.01
Star12P6	R	0, 1	1.40	-0.52	0.00	7673	0.54	0.55	-0.01	5431	0.76	0.75	0.01	1183	0.86	0.86	0.01
BANANTP6	R	R	0.82	0.29	0.03	7384	0.35	0.34	0.01	6666	0.56	0.55	0.01	1832	0.70	0.68	0.02
COUN10P6	R	R	1.25	-1.12	0.05	6661	0.82	0.79	0.03	6629	0.93	0.91	0.01	1836	0.97	0.96	0.01
COUN20P6	R	R	1.45	0.42	0.00	6661	0.20	0.24	-0.05	6629	0.56	0.51	0.05	1836	0.74	0.69	0.05
NUMBE4P6	R	R	2.81	-0.53	0.15	7965	0.61	0.61	-0.01	6769	0.89	0.88	0.02	1868	0.96	0.95	0.01
NUMBE7P6	R	0, 1	2.40	-0.28	0.07	7890	0.46	0.48	-0.01	5378	0.77	0.74	0.03	1186	0.89	0.87	0.03
NUMB17P6	R	0, 1	1.50	0.69	0.01	4854	0.22	0.23	-0.01	5261	0.33	0.30	0.03	1157	0.48	0.43	0.05
NUMB23P6	R	R	1.49	0.86	0.00	4850	0.13	0.17	-0.04	6581	0.38	0.35	0.03	1814	0.61	0.53	0.08
SQUAREP6	L	R	0.51	-1.96	0.10	2460	0.67	0.67	0.00	6823	0.89	0.88	0.01	1869	0.91	0.92	0.00
HouseP6	L	0, 1	0.78	-1.30	0.00	2509	0.51	0.50	0.00	5440	0.85	0.84	0.01	1189	0.89	0.89	-0.01
Finge4P6	L	0	1.21	-1.00	0.00	2463	0.41	0.38	0.03	1299	0.62	0.61	0.02	136	0.59	0.65	-0.06

See notes at end of table.

Table B-2. Mathematics item parameters and fit by wave—Continued

Item	Test Form PK	Test Form K06 & K07	IRT parameters			PK				K06				K07			
							P+				P+				P+		
			a	b	c	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff	N	Actual	Predicted	Diff
CRAYONP6	L	0, 1	0.77	-1.80	0.01	2500	0.67	0.65	0.02	5434	0.91	0.90	0.00	1187	0.95	0.94	0.02
5STICKP6	H	1, 2	1.56	1.08	0.20	864	0.55	0.50	0.05	4368	0.51	0.50	0.02	1478	0.59	0.61	-0.03
CatsP6	H	1	1.48	0.94	0.30	853	0.61	0.62	-0.01	3983	0.50	0.48	0.02	1014	0.55	0.54	0.00
EbonyP6	H	1, 2	0.95	1.01	0.15	850	0.56	0.53	0.03	5349	0.49	0.48	0.01	1705	0.57	0.58	-0.01
1plus7P6	H	2	2.69	1.56	0.32	825	0.42	0.39	0.03	1320	0.63	0.60	0.03	664	0.70	0.68	0.02
2plus2P6	H	2	3.24	1.25	0.18	563	0.43	0.40	0.03	1354	0.70	0.71	-0.01	675	0.80	0.80	0.01

† Not applicable.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool (2005–06), kindergarten 2006 (2006–07), and kindergarten 2007 (2007–08) data collections.