NCES 2018-182 U.S. DEPARTMENT OF EDUCATION

Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011)

Kindergarten Psychometric Report

NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011)

Kindergarten Psychometric Report

July 2018

Michelle Najarian
ETS

Karen Tourangeau
Christine Nord
Kathleen Wallner-Allen
Westat

Gail M. Mulligan
Project Officer
National Center for Education Statistics

NCES 2018-182
U.S. Department of Education

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

U.S. Department of Education
Betsy DeVos
Secretary

Institute of Education Sciences
Mark Schneider
Director

National Center for Education Statistics
James L. Woodworth
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

NCES, IES, U.S. Department of Education
Potomac Center Plaza (PCP)
550 12th Street, SW
Washington, DC 20202

July 2018

The NCES Home Page address is https://nces.ed.gov.
The NCES Publications and Products address is https://nces.ed.gov/pubsearch.

This publication is only available online. To download, view, and print the report as a PDF file, go to the NCES Publication and Products address shown above.

This report was prepared for the National Center for Education Statistics under Contract No. ED-04-CO-0059/0023 with Westat and ETS. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

Suggested Citation
Najarian, M., Tourangeau, K., Nord, C., and Wallner-Allen, K. (2018). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), Kindergarten Psychometric Report* (NCES 2018-182). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved [date] from http://nces.ed.gov/pubsearch.

Content Contact
Gail M. Mulligan
(202) 245-8413
Gail.Mulligan@ed.gov

CONTENTS

CONTENTS (continued)

CONTENTS (continued)

CONTENTS (continued)

CONTENTS (continued)

List of Appendixes

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF EXHIBITS

1. INTRODUCTION

This report describes the design, development, administration, quality control procedures, and psychometric characteristics of the child assessment instruments used to measure the knowledge, skills, and development of young children participating in the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) in the kindergarten year. The focus of this volume is the first two waves of data collection: the fall 2010 kindergarten wave and the spring 2011 kindergarten wave. This chapter provides a brief overview of the ECLS-K:2011 study; a discussion of the sample (section 1.1); an overview of the instrumentation (section 1.2); and an overview of the contents of this report (section 1.3).

The ECLS-K:2011 followed a nationally representative sample of students in U.S. schools from the time they were in kindergarten through their elementary school years. It is a multisource, multimethod study that focuses on the student's early school and home experiences. It includes interviews with parents, self-administered questionnaires completed by teachers and school administrators, and one-on-one direct assessments of students. During the kindergarten year, it also included self-administered questionnaires for nonparental before- and after-school care providers. The ECLS-K:2011 is sponsored by the National Center for Education Statistics (NCES) within the Institute of Education Sciences (IES) of the U.S. Department of Education.

The ECLS-K:2011 is the third and most recent study in the Early Childhood Longitudinal Study (ECLS) program, which comprises three longitudinal studies of young children: the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K); the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B); and the ECLS-K:2011. The ECLS program is unprecedented in its scope and coverage of child development, early learning, and school progress. It draws together information from multiple sources, including school administrators, parents, teachers, early care and education providers, and children, to provide data for researchers and policymakers to use to improve children's early educational experiences and address important policy questions. The ECLS-K:2011 provides current information about today's elementary school students and data relevant to emerging policy-related domains not measured fully in the previous ECLS studies. Also, coming more than a decade after the inception of the ECLS-K, the ECLS-K:2011 allows for cross-cohort comparisons of two nationally representative kindergarten classes experiencing different policy, educational, demographic, and economic environments.

Across the three studies, the ECLS program provides national data on children's developmental status at birth and at various points thereafter; children's transitions to nonparental care, early education programs, and school; and children's home and school experiences, growth, and learning. The ECLS program also provides data that enable researchers to analyze how a wide range of child, family, school, classroom, nonparental care and education provider, and community characteristics relate to children's development and to their experiences and success in school. Together these cohorts provide the range and breadth of data needed to more fully describe and understand children's educational experiences, early learning, development, and health in the late 1990s, 2000s, and 2010s.

More information about all three of these studies can be found on the ECLS website (https://nces.ed.gov/ecls).

1.1 The Kindergarten Student Sample

The ECLS-K:2011 provides national data on elementary school students' characteristics as the students progress from kindergarten in the 2010–11 school year through the spring of 2016 when most of them were in fifth grade. In the 2010–11 school year, the ECLS-K:2011 collected data on a nationally representative sample of about 18,170 students enrolled in approximately 970 schools across the United States.[1] During the kindergarten year there were two data collections, one at the beginning (fall) and one near the end (spring) of the school year.

The sample of students included in the ECLS-K:2011 was selected using a clustered, multistage probability design. In the first stage, 90 primary sampling units (PSUs), which are geographic areas made up of counties or groups of counties, were sampled. In the second stage, samples of public and private schools with kindergarten programs or that educated 5-year-olds in an ungraded setting were selected within the sampled PSUs. The third-stage sampling units were students enrolled in kindergarten and 5-year-olds in ungraded schools or classrooms who were selected within each sampled school. More detailed information about the sample design can be found in *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011),User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic Codebook, Public Version* (NCES 2015-074) (Tourangeau et al. 2015).

[1] The number of schools noted here is the number of schools that were sampled for participation in the study. It does not include schools to which sampled children transferred during the school year.

1.2 Data Collection Instruments and Administration of Assessments

The emphasis placed on measuring children's experiences within multiple contexts and development in multiple domains has critical implications for the design of the ECLS-K:2011. Data were collected on a wide array of topics at a broad level rather than on a select set of topics in more depth. Additionally, several different people from different contexts in the study child's life were asked to provide information about the child. Thus, the design of the study includes the collection of information from the students, their parents/guardians, their teachers, their schools, and their before- and after-school care (BASC) providers as described here:

Students were directly assessed in each round of the kindergarten year. The untimed assessments were administered to the sampled students, one-on-one, by a trained assessor. The students were assessed in reading (fall and spring), mathematics (fall and spring), and science (spring), as well as executive function[2] (fall and spring). The kindergarten year data collection also included an assessment of Spanish early reading skills (SERS) for Spanish-speaking English language learner (ELL) students who did not achieve a minimum score on an English language screener. In addition to the cognitive components, the assessments included measurements of the height and weight of all students in the fall and spring.

Parents/guardians were an important source of information about the study student, the student's family, and the student's home environment. Information was collected from parents in the fall and spring data collection rounds using computer-assisted interviews (CAIs). The parent interviews, most of which were conducted by phone,[3] asked about family structure, family literacy practices, parental involvement in school, nonparental care arrangements, household composition, family income, parent education level, and other demographic information. Parents were also asked to report on their children's health, socioemotional well-being, and disability status.

Teachers provided information about the students they taught, the students' learning environment at school, and themselves. More specifically, they were asked about their own backgrounds, teaching practices, and experience. They were also asked to provide information on the classroom experiences for the sampled students they taught and to evaluate each sampled student on a number of

[2] Executive functions are interdependent processes that work together to regulate and orchestrate cognition, emotion, and behavior and that help a student to learn in the classroom (e.g., Diamond 2013). More information about executive function is provided in Chapter 6.

[3] The parent interview was conducted in person when the parent did not have a telephone, was difficult to reach by telephone, or preferred to complete the interview in person.

critical cognitive and noncognitive dimensions. Information was collected from general classroom teachers via self-administered paper questionnaires during both the fall and spring data collections.

Special education teachers and service providers of sampled students who had an Individualized Education Program (IEP) in kindergarten were asked to provide information on the nature and types of services provided to the students, as well as on their own background and experience. Information was collected from special education teachers via self-administered paper questionnaires during spring data collection.

School administrators were asked to provide information on the physical, organizational, and fiscal characteristics of their schools, and on the schools' learning environment and programs. School administrators also provided information on their own background and experience. Information was collected from school administrators via self-administered paper questionnaires during spring data collection.

The kindergarten **before- and after-school care (BASC)** component collected important information about students' environments and experiences in nonparental care with regular before- or after-school care providers. Adults other than the student's parents/guardians who cared for the study student for at least 5 hours per week were asked to provide information such as the location where care was provided, children's activities while in care, characteristics of other children in care, and their own background and experience. Information was collected from care providers via self-administered paper questionnaires during spring data collection.

1.3 Contents of Report

This volume provides technical details about the design, development, and psychometric characteristics of the direct and indirect child assessments used during the fall 2010 and spring 2011 kindergarten waves of data collection. Chapter 2 provides details about the design of the direct child cognitive assessment battery for the ECLS-K:2011 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, including analysis of common item functioning. Chapter 4 discusses the development of the direct cognitive assessments through the field test item pool, item analysis, and results, including the development of the final

assessment forms and a description of item quality and reliability. Chapter 5 describes the psychometric characteristics of the direct cognitive assessment battery, including the approach to and types of scoring, choosing the appropriate scores for analysis, and measuring gains. Chapter 6 describes the psychometric characteristics of the executive function measures. Chapter 7 provides information on psychometric characteristics of the indirect measures, including the *Children's Behavior Questionnaire*, the social skills items adapted from the *Social Skills Rating System*, the *Approaches to Learning Scale*, and the *Student-Teacher Relationship Scale.* Following Chapter 7 are three appendixes supplementing the information in the main text. Appendix A provides basic information about the fall 2009 field tests conducted for the ECLS-K:2011, including an overview of the design of the survey instruments and assessments, and descriptions of the field tests, sample design, data collection and training, assessor feedback and recommendations, and analysis of the field test results. Appendix B provides details on how the students' responses to assessment items are prepared for and used in PARSCALE, the computer program used for estimating Item-Response Theory (IRT) models from which assessment scores are produced, as well as what quality control checks are performed on the assessment data. Appendix C lists the ECLS-K:2011 kindergarten reading, mathematics, and science IRT item parameters.

This page intentionally left blank.

2. OVERVIEW OF THE CONTENT COVERAGE AND ADMINISTRATION OF THE DIRECT COGNITIVE ASSESSMENT INSTRUMENTS

The direct cognitive assessments of reading, mathematics, and science developed for use in the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) were designed to meet several objectives within the design and scope of the study. First and foremost, the academic cognitive assessments needed to accurately measure children's acquisition of knowledge and skills throughout the elementary school years. The longitudinal design of the study required that a scale be developed in each subject area to support the measurement of change in knowledge and skills demonstrated by children from kindergarten entry through the spring of 2016 (when most students are expected to be in fifth grade), thus allowing for comparisons of achievement across grades and to quantify the gains children make over time. Also, as noted in chapter 1, a primary purpose of the ECLS-K:2011 is to enable comparisons between this kindergarten cohort and the cohort of students who were in kindergarten in the 1998–99 school year. In order for this comparison to be possible, there must be sufficient overlap in the content and actual items included in the assessments of the two studies. At the same time, the ECLS-K:2011 assessment needed to include new content and items reflecting differences or advancements in education policy, pedagogy, early childhood research, and society since the earlier study. Additionally, the goals of minimizing development and administration time, minimizing the cost of development and administration, and minimizing the burden on students and teachers affected the kinds of assessment items that could be used, as well as the structure of the assessments.

This chapter provides an overview of the academic cognitive assessments developed for use in the ECLS-K:2011, focusing on content and administration. Information about the assessment of executive function is provided in chapter 6, and information on the indirect measures of children's social skills, social relationships, and behavior problems is provided in chapter 7.

2.1 Content Coverage of the Cognitive Assessments

Child development and education experts were consulted by project staff during the design phase of the ECLS-K:2011 (see chapter 2 of *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), Kindergarten Methodology Report* for a listing of the experts). The experts recommended that the knowledge and skills assessed during each round of the ECLS-K:2011 should

represent the typical and important cognitive knowledge and skills covered by schools' curricula for the particular grade of interest. Also, the assessment content had to be similar to the content covered in the ECLS-K assessments. The ECLS-K assessed reading and mathematics starting in kindergarten, and the decision was made to do so in the ECLS-K:2011 as well. A general knowledge assessment administered in the kindergarten and first-grade data collections of the ECLS-K included some science items, but a separate science assessment was not fielded until the third-grade data collection. The ECLS-K assessment schedule makes it possible to analyze gains in general knowledge (which included science) from the fall of kindergarten to the spring of first grade, and to analyze gains in science knowledge and skills from third grade through eighth grade, but it is not possible to measure gains in science knowledge and skills from the beginning to the end of that study. This limitation in the assessment design, along with a strong interest in including a separate science assessment in the ECLS-K:2011 as early as possible from expert consultants, led to the decision to include a separate science-only assessment beginning in the kindergarten year of the ECLS-K:2011. Thus, in the ECLS-K:2011, it will be possible to analyze children's science knowledge and skills throughout the study, from the spring of kindergarten to the last round of data collection in the spring of 2016.

In addition to providing a consistent measure of science knowledge over the course of the elementary school years, administering a science-only assessment in all grades included in the ECLS-K:2011 is consistent with more recent emphasis on STEM subjects (science, technology, engineering, and mathematics) in education policy and K-12 education. For example, in February 2006, the Deficit Reduction Act called for the establishment of the Academic Competitiveness Council (ACC) to identify, evaluate, and recommend improvements to federal programs with a mathematics or science education focus. The year-long ACC study, analyzing data from several sources, including results of National Assessment of Educational Progress (NAEP) mathematics and science assessments, found that there is increasing concern about the United States' economic competitiveness, particularly the future ability of the nation's educational institutions to produce citizens literate in STEM concepts and to produce future scientists, engineers, mathematicians, and technologists (U.S. Department of Education 2007). Also, the Elementary and Secondary Education Act requires that, beginning in 2007, states measure students' progress in science at least once in each of three grade spans (3–5, 6–9, 10–12) (U.S. Department of Education 2004). Many states across the United States have responded to this mandate by developing science standards for each grade level from kindergarten through 12th grade. The science assessments in the ECLS-K:2011 will provide information about variation in elementary school students' science knowledge and skills at the national level.

The following sections describe the frameworks that guided the development of the assessment for each ECLS-K:2011 assessment subject area (reading, mathematics, and science) from kindergarten through third grade.[1] Supporting information from current curriculum standards for each subject area is also provided.

2.1.1 Reading Test Specifications

The content category specifications for the ECLS-K:2011 reading assessment are largely based on the 2009 Reading Frameworks for NAEP (National Assessment Governing Board 2008). Although the NAEP framework was selected for its rigorous design and its use in many years of national administrations by NCES, because the NAEP assessments are administered starting in fourth grade, it was necessary to consult other sources to extend the NAEP content percentage specifications down to the kindergarten level. In addition to the ECLS-K kindergarten, first-grade, and third-grade reading assessment frameworks,[2] current curriculum standards for kindergarten through third grade from California, Florida, New Jersey, Texas, and Virginia[3] were consulted by experts in reading assessment development. The experts agreed that the ECLS-K frameworks, which included the addition of a basic reading skills category to the content areas included in NAEP, as well as the inclusion of grade-appropriate vocabulary items, would ensure that the ECLS-K:2011 assessments would be suitable for kindergarten through third grade. Thus, the framework that guided the development of the ECLS-K:2011 reading assessment includes items measuring knowledge and skills in three broad categories: basic skills, vocabulary, and reading comprehension.

2.1.1.1 Basic Reading Skills

Basic reading skills include many early literacy skills such as phonological awareness, familiarity with print, recognition of letters and sounds, and identification of common sight words.

[1]Although this report focuses on the kindergarten assessments, the longitudinal nature of the study and the need to be able to measure gains over time required the development of a framework spanning multiple grades before the assessment for any one grade could be developed. The discussion about the overall framework from kindergarten through third grade is included here as a reference for what content is to be expected in the subsequent rounds, as well as how the kindergarten content relates both to the overall framework and to the content measured in later rounds.

[2] There was no framework or national administration of a second-grade assessment in the ECLS-K.

[3] These states were selected because they span the contiguous United States and the curriculum standards from these states were familiar to the assessment developers, who had extensive experience in item development for assessments for these states.

Phonological awareness is one of the major reading skills included in the ECLS-K:2011 assessments. Phonological awareness is a broad term used to describe the manipulation of spoken word parts, including phonemes, syllables, onsets, and rhymes. The acquisition of phonemic awareness is highly correlated with success in reading. Studies show that these skills also aid in reading comprehension (International Reading Association, 1998). To become more fluent readers, many children rely on their decoding skills. Decoding is the ability to apply knowledge of letter-sound relationships in order to read unknown words. Students who are taught phonemic awareness have skills enabling them to read unfamiliar words quickly and accurately. Once decoding is mastered, reading fluency becomes much easier. Readers are then able to develop further their comprehension skills by focusing their attention on the meaning of texts (Adler 2003). Phonological awareness was evaluated in increasing difficulty, beginning with broad skills and advancing to more specific skills (Vukelich and Christie 2004). Specifically, the ECLS-K:2011 reading assessment measures the following types of phonological awareness skills:

- Rhyming (e.g., naming words that rhyme with a stimulus word);
- Sound matching (e.g., pointing to a picture showing something that begins with the same sound as the stimulus picture, for example, a sock and sand);
- Initial and final sounds of words (e.g., pointing to the letter that makes the same sound heard at the beginning or end of a stimulus word);
- Blending (combining sounds to form a word);
- Segmentation (identifying the number of sounds in a word); and
- Manipulation of phonemes (adding, deleting, or substituting sounds, for example, asking what the new word would be if a new sound was added to the end of a stimulus word or if the first sound of a stimulus word was replaced with a different first sound).

Familiarity with print refers to children's understanding of the way text is structured (for example, knowing that in English, text is read from left to right), and how it is used to convey meaning. Skills and knowledge such as demonstrating an understanding of the concept of "a word" or "a sentence," knowing the difference between text and illustrations, and understanding the use of punctuation are also valuable to understanding the structure of text. Assessment tasks such as having a child demonstrate how to hold a book correctly, asking where the cover of the book is, where the title of the book is, how to turn pages in a book, and how text is read (from left to right, top to bottom) can show a child's knowledge of print conventions (International Reading Association [IRA] and the National Association for the Education of Young Children [NAEYC], 2008). The ECLS-K:2011 assessment includes several items

like these, for example, asking children where the assessor should start to read if the assessor wants to read the text on a page of a book.

Recognition of letters and sounds connects spoken language to written language. This is one of the first skills in early reading (IRA, and NAEYC 2008). ECLS-K:2011 assessment tasks related to letter and sound recognition ask children to perform tasks, such as choose a specific letter from a set or to give the name of a letter that is shown to them. These tasks involve identification of both upper and lower case letters. In addition, children are asked to associate a letter with its sound. These tasks include the child identifying the letter that makes a sound vocalized by the assessor or the child vocalizing the sound represented by a certain letter named by the assessor.

Sight words are high-frequency words children are likely to encounter every day. Recognizing sight words easily and quickly enables children to become more fluent readers. The ECLS-K:2011 assessment measures children's knowledge of sight words of varying difficulty taken from the Dolch sight word list (Dolch 1948).

2.1.1.2 Vocabulary

Vocabulary knowledge represents understanding of the meanings of words. Although children may be able to decode printed text, they also must understand the meaning of the words they have read in order to be able to comprehend the text. Vocabulary test questions in the ECLS-K:2011 assessment ask children to convey their vocabulary knowledge both verbally (expressive vocabulary) and nonverbally (receptive vocabulary). With expressive vocabulary, a child is asked a question associated with a stimulus picture, for example, "What is this?" and gives a verbal response. With receptive vocabulary, a child is given a vocabulary word and asked to select a certain object representing that word from a group of similar objects using a nonverbal response (e.g., pointing to a picture presented in the assessment easel). This task assesses the child's understanding of the given word in relation to a picture of it. In addition, some receptive vocabulary tasks pertain to words used in context and assess the reader's ability to use the text as an aid for clarifying the meaning of unfamiliar words. Children are given a word in the context of a sentence or paragraph and asked to identify a word or phrase that means the same thing. Because this task requires children to be able to read, it measures vocabulary knowledge at a deeper level than asking them to point to the picture representing a stimulus word.

2.1.1.3 Comprehension

As noted earlier, the ECLS-K:2011 reading framework was modeled after the NAEP 2009 reading framework. The **locate/recall, integrate/interpret,** and **critique/evaluate** content categories, which were derived directly from the NAEP framework, measure children's reading comprehension skills and rely on their ability to read text independently (National Assessment Governing Board 2008).

- **Locate/recall.** Test questions in this category ask readers to identify information explicitly stated in the text, such as definitions, facts, and supporting details, and to make simple inferences within and between texts. For example, a child is asked to give the name of a pet featured in a story the child read or to list the three things that fell out of the backpack of a girl featured in a different story.

- **Integrate/interpret.** Test questions in this category ask readers to make complex inferences within and between texts to describe a problem and solution, or cause and effect. Questions assess the child's ability to go beyond the text to arrive at a logical conclusion. Questions in this category also ask the child to summarize ideas, draw conclusions, or predict outcomes. For example, a child is asked why the two characters in a story are friends after reading about how one of the characters helped the other.

- **Critique/evaluate.** Test questions in this category ask readers to consider texts critically by asking them to consider the text objectively and judge its appropriateness and quality. These types of questions provide information on critical skills as early as kindergarten. For example, a child is asked what information about a missing pet would be helpful for people looking for the pet to know.

These reading comprehension skills are assessed in the ECLS-K:2011 assessments by having children read various literary and informational texts, and then asking them questions about what they read. Reading literary text in elementary school involves exploring themes, characters, events, problems, and settings of literary works in a variety of genres, including stories, poetry, plays, myths and legends, and novels. Reading for information in elementary school involves relating the information in the text with aspects of the real world and is most commonly associated with reading textbooks and newspaper and magazine articles. All of the passages in the kindergarten assessment and most of the passages in the first-grade assessment are literary texts. The number of informational texts and their level of sophistication increases gradually in the ECLS-K:2011 testing battery, such that in second and third grade, approximately two-thirds of the passages will be literary texts and one-third of the passages will be informational texts. It should be noted that, even though a relatively small proportion of kindergartners can read for understanding, reading passages and associated reading comprehension questions are included in the kindergarten assessment to ensure that the knowledge and skills of the highest performing

children are adequately measured and to include the higher-difficulty items that are needed to enable the linking of assessments across grades.

2.1.1.4 Continuity Between the ECLS-K and the ECLS-K:2011 Reading Frameworks

Continuity between the ECLS-K and ECLS-K:2011 framework specifications was necessary to develop an ECLS-K:2011 reading assessment measuring similar content as the ECLS-K reading assessment to enable cross-cohort comparisons, which is one of the goals of the ECLS-K:2011. The content categories of the ECLS-K reading assessment framework, which was modeled after the 1992 and 1994 NAEP frameworks (National Assessment Governing Board 1993; National Assessment Governing Board 1996), correspond to the ECLS-K:2011 reading framework content categories. The **basic skills** and **vocabulary** categories are similar in both the ECLS-K and the ECLS-K:2011 frameworks. The ECLS-K category **forming a general understanding** closely corresponds to the ECLS-K:2011 **locate/recall** category. The ECLS-K:2011 **integrate/interpret** category combines the **developing interpretation** and the **making reader-text connections** categories of the ECLS-K. The **examining content and structure** category of the ECLS-K is similar to the ECLS-K:2011 **critique/evaluate** category.

2.1.2 Mathematics Test Specifications

The mathematics test specifications for the ECLS-K:2011 are based primarily on the frameworks developed for the ECLS-K kindergarten, first-grade, and third-grade mathematics assessments. The ECLS-K framework was based on the NAEP 1996 mathematics framework (National Assessment Governing Board 1994) and extended down to earlier grades (Rock and Pollack 2002; Pollack et al. 2005). For second grade, the ECLS-K:2011 framework could not be based on that from the ECLS-K, since there was no national administration of a second-grade assessment in the ECLS-K. A review of current state curriculum standards suggested that the skills covered in second grade closely match those taught in first grade. Consequently, the ECLS-K:2011 mathematics framework for second grade is closely aligned with that for first grade.

When the ECLS-K:2011 mathematics framework was being developed, the 2005 NAEP fourth-grade mathematics framework (Lee, Grigg, and Dion 2007) was reviewed and found to have changed only minimally from the 1996 framework. Given this, along with the need to have continuity

between the ECLS-K and the ECLS-K:2011 assessment frameworks to allow for cross-cohort comparisons, the decision was made to use the ECLS-K framework as the basis for the ECLS-K:2011 mathematics assessment, rather than use a more recent version of the NAEP framework as was done for reading. However, even though the ECLS-K:2011 mathematics framework is based on older specifications, the final content of the mathematics framework is consistent with recommendations presented in the Mathematics Framework for the 2005 NAEP (National Assessment Governing Board 2004); the National Council of Teachers of Mathematics *Principles and Standards for School Mathematics* (2001); and with state standards of California, New Jersey, Tennessee, Texas, and Virginia.[4] The content is also consistent with general recommendations from the National Mathematics Advisory Panel (2008). The framework that guided the development of the ECLS-K:2011 mathematics assessment includes the content categories: **number properties and operations**, **measurement, geometry, data analysis and probability,** and **algebra**.

2.1.2.1 Number Properties and Operations

In grades K–3, this content area largely assesses number sense, which refers to children's understanding of numbers, operations, and estimation and their application to real-world situations. Number sense also involves being able to read and write numbers and having an understanding of mathematics language and symbols. At the kindergarten level, students may be developing an awareness and ability to match number words with the appropriate numeral, and finding sums or differences using numbers less than 20, when given concrete models or pictures. As children advance in age and grade, they will be required to expand the foundation of knowledge to building a system of tens; using larger numbers; applying operations to larger numbers; ordering and comparing whole numbers, fractions, or decimals; and applying mathematical ideas to real-world situations. Additionally, children will be required to move from concrete representations of operations and ideas to more abstract representations and algorithms. In the kindergarten assessment, this content category is measured with questions asking children to identify certain single-digit and two-digit numbers; simple addition, subtraction, multiplication, and division problems using numbers less than 20; and items assessing knowledge of relative quantity (e.g., more than, less than, or equal to).

[4] These states were selected because they span the contiguous United States and the curriculum standards from these states were familiar to the assessment developers, who had extensive experience in item development for assessments for these states.

2.1.2.2 Measurement

Measuring is the process by which numbers are assigned in order to describe the world quantitatively. Measurement skills include choosing a measurement unit, comparing the unit to the measured object, and reporting the results of a measurement task. This content area includes items assessing children's understanding of how to measure using standard and nonstandard units and the concepts of time, money, temperature, length, perimeter, area, mass, and weight. In kindergarten, students should be able to compare objects by attribute and tell general times of the day (day, night). As children advance in age and grade, they should be able to use measurement tools to measure time, temperature, length, mass, and weight and later extend into more advanced concepts such as perimeter, area, and volume. In the kindergarten assessment, this content category is measured with questions asking children to identify objects that are shorter than a stimulus object, to indicate the length of an object as measured on a ruler, and to perform basic operations that require knowledge of money.

2.1.2.3 Geometry

In this content area, students are expected to be familiar with geometric figures and their attributes, both in the plane (lines, circles, triangles, rectangles, and squares) and in space (cubes, spheres, and cylinders). In kindergarten, children are expected to identify only simple plane shapes such as triangles, circles, and squares. As children advance in age and grade, they should expand their knowledge into other plane shapes and three-dimensional figures, including polygons and polyhedrons, and determine the results of putting together and taking apart two- and three-dimensional figures. In the kindergarten assessment, this content category is measured with questions asking children to identify basic shapes.

2.1.2.4 Data Analysis and Probability

Data analysis covers the entire process of collecting, organizing, reading, representing, and interpreting data. Children in kindergarten are asked to compare or draw simple conclusions about a set of data while older children may be asked to identify patterns, make inferences, or draw conclusions based on the data. Probability refers to making judgments about the likelihood of something occurring. Children in kindergarten are asked if something is more or less likely to occur, while older children may be asked

to give a numerical probability of an outcome given a set of data. In the kindergarten assessment, this content category is measured with questions asking children to read basic graphs and the probability of coins landing heads up.

2.1.2.5 Algebra

Algebra refers to the techniques of identifying solutions to equations with one or more missing pieces or variables and completing patterns. Specifically, children are evaluated on their ability to recognize, create, explain, generalize, and extend patterns and sequences. In the kindergarten assessment, this content category is measured with questions asking children to complete patterns involving numbers and patterns involving shapes. As children advance in age and grade, algebraic equations and functions will be added.

2.1.3 Science Test Specifications

The science knowledge and skills assessed in the ECLS-K:2011 were chosen based on the areas identified as being important to assess in the 1996–2005 NAEP science frameworks. They encompass the knowledge and use of organized factual information; understanding of the relationships among Earth, life, and physical science concepts; major ideas unifying the different areas of science (e.g., chemistry, biology); and thinking and laboratory skills (National Assessment Governing Board 2004b). However, because the NAEP frameworks begin in fourth grade, the standards of six states (Arizona, California, Florida, New Mexico, Texas, and Virginia)[5] were analyzed to find a commonality of topics that are taught at the lower grade levels assessed in the ECLS-K:2011. Across these states and for each grade level, three or four standards were specified for each of four common reporting categories: **scientific inquiry, life science, physical science, and Earth and space science.** These four reporting categories were selected as the content categories for the ECLS-K:2011 science assessment framework.

[5] These states were selected because they span the contiguous United States and the curriculum standards from these states were familiar to the assessment developers, who had extensive experience in item development for assessments for these states.

2.1.3.1 Scientific Inquiry

In this content area, children in kindergarten are expected to observe common objects using the five senses, describe the properties of common objects by direct observation, sort common objects by physical attributes, and record observations and data. In subsequent grades, children are expected to collect information using measurement tools (e.g., clocks, thermometers), draw inferences and conclusions about familiar objects and events, conduct simple investigations, predict the outcome of a simple investigation, and compare results with the predictions. In the kindergarten assessment, this content category is measured with an item about a microscope and one about the properties of a rock.

2.1.3.2 Physical Science

In kindergarten, children are expected to make observations that different materials have different properties and that objects are made of different types of materials; compare the relative sizes and characteristics of objects; and investigate the way things move and observe differences. In subsequent grades, children are expected to identify the three states of matter; observe the different ways things may move; observe the effects of electrically charged materials and magnets; understand the basic properties of solids, liquids, and gasses; and understand that energy comes from the Sun to the Earth in the form of light and heat. In the kindergarten assessment, this content category is measured with questions about energy and about the materials from which common objects such as a nail or paper clip are made.

2.1.3.3 Life Science

In kindergarten, children are expected to recognize the five senses and the related body parts, identify major structures and functions of parts of plants and animals, and describe the similarities and differences in the appearance and behavior of plants and animals. In subsequent grades, children are expected to understand that living organisms inhabit various environments, understand how the environment influences some characteristics of living organisms, know that plants and animals have structures and adaptations that serve different functions, and know specific details about the life cycle of plants, including the fact that roots are associated with the intake of water and soil nutrients and that green leaves are associated with making food from sunlight. In the kindergarten assessment, this content

category is measured with questions related to the function of the human stomach, which insect makes honey, the way animals move and what they eat, and the development of a tadpole into a frog.

2.1.3.4 Earth and Space Science

In kindergarten, children are expected to observe that changes in weather occur from day to day and season to season; identify patterns in nature; and describe properties of rocks, soil, and water. In subsequent grades, children are expected to understand how weather affects people's daily activities, understand that shadows are caused when sunlight is blocked by objects, know the relationship between the Sun and the Earth, understand the processes involved with soil formation, be familiar with the processes in the water cycle, understand the movement of the Sun, Moon, and stars, and understand the relationship of objects within the solar system. In the kindergarten assessment, this content category is measured with weather-related questions, such as what the weather is like at the South Pole, questions about animal habitat, and an item asking about the best way to describe a rock shown in a stimulus picture.

2.2 Item and Time Allocation Across and Within Subject Areas

For all rounds of data collection, the overall testing time for each child was expected to be approximately 60 minutes, with more time allotted for the reading assessment (30 minutes) than for the mathematics (17 minutes) and science (13 minutes) assessments. A primary reason for this difference in overall timing across subject areas is that the reading assessment includes passages that need to be read before questions assessing knowledge and skills can be asked. Many mathematical and science items can be administered in a short period of time, while reading questions based on passage comprehension require a greater investment of time.

As stated above, the relative emphasis given to different content categories within each subject area assessment reflects the typical curriculum emphases. The general rule used in determining the item content allocations was that the composition of the tests should reflect the main content areas covered by the curriculum for each grade while simultaneously considering differences in the number of items and length of time needed to complete the items in order to adequately measure a given skill, knowledge, or concept. Systematically collected evidence on typical curricular content is not available in

most subject areas, so the study relied mainly on the advice of curriculum specialists and experts with extensive teaching and administrative experience in schools and on the standards published by states and national professional organizations.

In addition to the content categories, the specifications for the ECLS-K:2011 assessments in each subject area further indicate the approximate percentage of the items in the assessment for each grade level that fall within each of the content categories. The distribution of items in the reading assessment by content category and grade level is summarized in table 2-1 as target percentages of items. Assessments in the lower grades typically contain more items from content categories that are, in general, easier (e.g., letter identification in the basic skills content area), while assessments in the higher grades typically contain more items from more difficult content categories (e.g., recalling information in a reading passage). This can be seen in the pattern of percentages in the table, for example, where the percentage of items in the basic skills category decreases from kindergarten to third grade while the percentage of items in the critique/evaluate category increases. In order to adequately capture variation in the knowledge and skills of younger students who are just learning to read, the assessment needed to have a relatively larger proportion of items measuring basic skills and vocabulary acquisition. The percentages in kindergarten and first grade are heavily weighted toward those two categories for this reason. In contrast, a larger percentage of the items in the assessments for older students (second- and third-graders) who have begun to read and whose reading comprehension is increasing, assess skills that are more cognitively complex. Note that while in some instances the framework percentages (in reading, mathematics, and science) remain the same across grade levels, the assessments do not. For example, although vocabulary items account for 15 percent of the overall items in the kindergarten and first-grade reading assessments, in kindergarten the vocabulary items administered are, on average, less difficult than those administered in first grade.

Table 2-1. Reading content categories and target percentages of items, ECLS-K:2011 kindergarten through third grade assessments

Grade level	Basic skills	Vocabulary	Locate/recall	Integrate/interpret	Critique/evaluate
Kindergarten	50	15	20	10	5
1	40	15	20	20	5
2	20	10	30	30	10
3	15	10	30	30	15

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), provisional cognitive assessment frameworks, fall 2008.

The distribution of items in the mathematics assessment by content category and grade level is summarized in table 2-2 as target percentages of items. Similar to reading, assessments in the lower grades typically contain more items from content categories that are, in general, easier (e.g., number identification in the number properties and operations content area), while items in the higher grades typically contain more items from more difficult content categories (e.g., algebra skills). This can be seen in the pattern of percentages in the table, for example, where items in the number properties and operations content category constitute 75 percent of the content in the assessments in kindergarten through second grade, with the remaining 25 percent of items distributed across the four other content areas. There is a large shift in third grade toward a lower percentage of items in the number properties and operations category with a concurrent increase in the percentage of items in the other four content areas.

Table 2-2. Mathematics content categories and target percentages of items, ECLS-K:2011 kindergarten through third grade assessments

Grade level	Number properties and operations	Measurement	Geometry	Data analysis and probability	Algebra
Kindergarten	75	5	3	8	9
1	75	5	3	8	9
2	75	5	3	8	9
3	40	20	15	10	15

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), provisional cognitive assessment frameworks, fall 2008.

The distribution of items in the science assessment by content category and grade level is summarized in table 2-3 as target percentages of items. Unlike in the reading and mathematics assessments, the percentage of items for each science content category in each grade level is the same so that no category is overrepresented in the assessment. This follows common practice among states to represent each of these content strands equally within their curriculum standards.

Table 2-3. Science content categories and target percentages of items, ECLS-K:2011 kindergarten through third grade assessments

Grade level	Scientific inquiry	Life science	Physical science	Earth and space science
Kindergarten	25	25	25	25
1	25	25	25	25
2	25	25	25	25
3	25	25	25	25

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), provisional cognitive assessment frameworks, fall 2008.

2.3 Mode of Administration

The ECLS-K:2011 implemented many of the well-tested procedures developed for and used throughout multiple rounds of data collection in the ECLS-K and the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B). One of these procedures is to administer the assessment to each student individually. Since young children are generally not experienced test takers, individual administration by a trained assessor allows for more sensitivity to each child's needs than does a group-administered test. Also, children's performance during individual administration is more likely to reflect their true knowledge and skills as opposed to their test-taking proficiency.

Assessors used computer-administered personal interview (CAPI) technology to administer the assessments. With CAPI, the computer prompts the assessor to administer the items using a visual stimulus shown to the children in a spiral-bound book called an *easel*. For each assessment item, the CAPI program also provides the assessor with a standardized administration protocol, the question to be read verbatim to the child, and any instructions that should be provided to the child. Assessors entered all of the children's responses into the CAPI program.

In addition to being individually administered, the reading and mathematics assessments were also adaptive in nature, similar to the assessments in the earlier ECLS studies; that is, each child was administered a set of items that was most appropriate for that child's level of knowledge and skills. This procedure reduced the time burden on children, as well as the likelihood that children would become frustrated by being asked questions that were too easy or too difficult for them.

Psychometrically, adaptive tests are in general significantly more efficient than "one form fits all" administrations. Two-stage adaptive testing uses performance at the beginning of a testing session to direct the selection of later tasks that are at an appropriate difficulty level for each child. The reliability per unit of testing time is greater than it is when one standard form is used (Lord 1980). Adaptive testing also reduces the potential for floor and ceiling effects, which can affect the measurement of gain in longitudinal studies. Floor effects occur when some children's ability level is below the minimum that is accurately measured by a test. This can prevent low-performing children from demonstrating their true gains in knowledge when they are retested. Similarly, ceiling effects result in failure to measure the gains in achievement of high-performing children whose abilities are beyond the most difficult test questions.

In fully adaptive computerized testing, the selection of every item administered to a test taker is determined during the test and is based on the test taker's responses to the questions already answered. Fully adaptive computerized testing is not operationally feasible for the ECLS-K:2011, given the format of the assessment. The reading and mathematics components of the ECLS-K:2011 kindergarten assessment battery were multistage adaptive tests. The science and the Spanish early reading skills (SERS) assessments were single-stage only. In the multistage adaptive assessments, all children were first administered a routing test with items that varied in level of difficulty. Assessors entered children's responses into the CAPI program, which calculated a score for the child on the routing test. The child's score on the routing test determined which one of three second-stage tests (low, middle, or high difficulty) the child was administered. Thus, the test is adaptive in that children are administered *groups* of items based on their demonstrated performance on the routing test.

Although the second-stage tests are tailored for particular ability level ranges within a grade, the *overall* assessment reflects core curriculum elements for the particular grade being tested. Thus, a child who is essentially performing on grade level receives items that span the curriculum for that child's grade. Children whose achievement is above or below grade level are given items with difficulty levels that match their individual level of knowledge and skills at the time of testing rather than a grade-level standard. A child who is performing much better in relation to his or her peers, as measured by the routing test, is given test items that are proportionately more difficult (including some above grade level), while a child performing below grade level receives a second-stage test with proportionately more easy items (including some below grade level).

As noted earlier, two cognitive assessments administered in the kindergarten data collections were not adaptive. The science assessment, which was administered only in the spring kindergarten data

collection, was a single-stage test. Through analysis of the field test data, it was determined that children's abilities in science were not as diverse as originally anticipated; thus, a single-stage science assessment in kindergarten was deemed adequate. More information on the design of the science assessment can be found in section 4.3.3.3. A single-stage Spanish early reading skills (SERS) assessment was administered in kindergarten to Spanish-speaking children who lacked a sufficient level of English proficiency to proceed with the full assessment battery in English. More information about the English proficiency assessments used in the kindergarten rounds and routing of children through the cognitive assessment battery can be found in sections 4.3 and 5.1.

2.4 Inclusion of Children With Diverse Language Backgrounds and Language of Administration

The assessment procedures developed for the ECLS-K:2011 need to accommodate children with diverse language backgrounds. While the majority of the children in the study speak English as their first and only language, many of them speak a language other than English at home. Some of the children in the latter group also speak English at home while others do not. Because the educational environment in most U.S. schools is English dominant, and it is cost prohibitive to develop fully comparable assessments in different languages, the ECLS-K:2011 assessments were primarily administered in English. However, several of the assessments were translated into and administered in Spanish.

The routing within the ECLS-K:2011 cognitive assessment for children with a non-English primary home language differed somewhat from the ECLS-K. In the ECLS-K, children with a non-English primary home language who did not achieve at least a minimum score on an English language proficiency screener (referred to as the Oral Language Development Scale, or OLDS) were not administered any of the cognitive assessments in English. Spanish-speakers who did not pass the OLDS were administered a Spanish version of the mathematics assessment. The reading and general knowledge assessments were not translated into Spanish in the ECLS-K. This assessment procedure was instituted for the benefit of the children, so that they would not be administered an assessment in a language they did not understand. However, it was disadvantageous from an analytic perspective, because it is only possible to assess those children's gains in reading skills and general knowledge beginning with the first round in which they passed the language screener and took the assessments in English.

To address this limitation in the assessment design of the ECLS-K, an English reading assessment was developed for the ECLS-K:2011 that measures all children's English reading knowledge and skills, regardless of home language, from the first round in which children are assessed. An English language proficiency screener was used in the ECLS-K:2011, but unlike the procedures used in the ECLS-K, results from the screener were used to route children whose primary home language was not English (according to school report) out of the cognitive assessments in English *after* being administered a subset of the items in the full reading assessment, rather than routing them out of the cognitive assessments in English completely. Specifically, all children were administered the first 20 items of the full reading assessment, regardless of their English proficiency; these items provide information on children's basic reading skills in English (and thus are referred to collectively as the EBRS for "English basic reading skills") and are sufficient to compute an English reading assessment score. The EBRS contains items from the first two reading content categories, basic reading skills and vocabulary, and measures skills such as letter recognition, letter sounds, phonemic awareness, beginning and ending word sounds, and one sight word.

After completing the EBRS, Spanish-speaking children who did not achieve at least a minimum score on the language proficiency screener were administered an assessment of reading knowledge and skills in Spanish (described further in the next paragraph), as well as the mathematics and executive function assessments translated into Spanish. The science assessment was not translated into Spanish. They also had their height and weight measured. Children whose primary home language was not English or Spanish, who did not achieve at least a minimum score on the language proficiency screener, only had their height and weight measured after completing the EBRS.

Research on language acquisition suggests that skills in one language can be transferred to another language. As a result, children who are proficient in one language can learn a new language more easily, compared to children who begin to learn a new language without having a solid foundation in at least one language (Odlin 1989). Although children whose primary home language is not English may not have developed reading knowledge and skills in English before entering kindergarten, they may be developing such skills in their home language. In order to assess the development of early reading skills in general, regardless of the language in which they are being developed, a Spanish early reading skills (SERS) assessment, which includes items that measure letter identification, letter sounds, print familiarity, and simple vocabulary, was administered in the ECLS-K:2011 to children who were not proficient in English and for whom the school reported a home language of Spanish. The SERS items are translations of the items also fielded as part of the English reading assessment in the EBRS and are not

intended to be a measure of *proficiency* in Spanish. Rather, results from the SERS are intended to provide additional information about Spanish-speaking children's basic language skills that may be predictive of their success in school. More information on the development of the assessments in Spanish and the scores available for them can be found in chapters 4 and 5, respectively.

This page intentionally left blank.

3. ANALYSIS METHODOLOGY

This chapter describes the standard procedures used to process data from the ECLS-K:2011 direct child cognitive assessments, both for item selection (using field test data) and to produce scores for analysis (using national administration data). It provides background for understanding the results discussed in chapters 4 and 5. The chapter begins with a brief discussion in section 3.1 of the quality control steps followed in order to ensure that the data used for scoring were accurate. An overview of the item response theory (IRT) model and procedures (Lord 1980), used to carry out psychometric analysis of the data, is provided in section 3.2. IRT methodology is used to put scores that are obtained on different sets of test items on the same scale for comparison within and across assessment years. In addition to scoring, the selection of the IRT model, evaluation of items using empirical item characteristic curves, and item information and measurement precision are discussed, followed by a brief explanation of IRT using PARSCALE. Section 3.3 discusses the examination of dimensionality in order to assess the construct validity of the assessments, followed by section 3.4 with the details of procedures used to examine differential item functioning (DIF), which identify test items that perform differently for certain subgroups of the population when ability is held constant. Section 3.5 discusses the evaluation of common items and the development of the longitudinal scales that are produced using assessment data from both the fall and spring of kindergarten, which allow for the measurement of gains in knowledge and skills across time.

3.1 Quality Control Procedures

Many procedures were employed to ensure that the data used to produce the reading, mathematics, and science assessment scores were accurate and valid. Before data collection began, the programming for the computer-assisted personal interview (CAPI) system was reviewed and tested to ensure that the system was accurately capturing child responses to the assessment items, calculating correct scores for the first-stage routing test, and routing the child to the appropriate second-stage test based on the routing score. After the data collection ended, during the process of estimating final scores from the raw item response data, response frequencies were reviewed for each item, item functioning was evaluated using both classical item analysis and IRT methods, and the item data were assembled into meaningful and interpretable scores.

For each round of data collection, frequency distributions of raw item responses were produced for each test item to serve as a baseline for confirming the accuracy of later processing steps. Each distribution of responses was compared with the text of the corresponding question in the assessment and with the instructions the assessor used when recording responses to confirm that responses were coded accurately. For example, the distribution of responses for a four-option multiple-choice question would be expected to contain response codes of 1, 2, 3, and 4. Responses of 1 (correct) or 2 (incorrect) were to have been recorded by the assessors for dichotomous open-ended questions (i.e., those without predetermined response options from which to choose). Missing data codes (-8 = refused, -9 = don't know, blank = not administered) also were counted for each item.

Classical item analysis, which 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), was carried out separately for each round of data collection and for each subject area assessment using Educational Testing Service's (ETS's) proprietary software, F4STAT. Sets of statistics were produced for each item, as well as summary statistics for each individual form. Each of these statistics provided information on item performance and was an additional 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. Additionally, the same statistics were computed separately for items identified as "omits" and for items identified as "not reached." "Omits" are items children refused to respond to or multiple-choice items for which they responded "don't know," that were followed by at least one subsequent item the test taker did answer. "Not reached" items are those for which test takers provided no answer and for which no subsequent item had a response, which could occur when an assessment was discontinued due to burden on the test taker or refusal by the test taker to continue. 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 as "omits" or "not reached" correctly.

Summary statistics produced for each item included the proportion correct and *r*-biserial. The *r*-biserial is the correlation of the item score (i.e., whether it was correct or not) with the total number-right score for its test section (for example, the router or the low-, middle-, or high-level second-stage test), 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 in the output, 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 responses, and the *r*-biserial should be positive, ideally .30 or higher. If these conditions are not satisfied, one of two error conditions could be responsible. The correct answer may not have been correctly identified, 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 occur for an item that is much too easy or much too hard for the vast majority of test takers. If virtually all test takers could answer an item correctly or, at the opposite extreme, virtually all could only guess at the answer, the variance in item score (i.e., whether the item was answered correctly) would be low or nonexistent. Consequently, the resulting correlation of the percent correct for the item with total test score (adjusted to compute the *r*-biserial) also would be low. Summaries, by domain, of the classical item analysis results can be found in chapter 5.

During test development (which is described in chapter 4), items within each test section or group of items of the same content type were arranged in ascending order of anticipated difficulty, based on results from the field test analyses. A review of an item's percent correct statistics allows for the identification of any serious deviation from this expectation, which could indicate anomalies in the administration or scoring of the item. 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 (see section 4.3.3).

Summary statistics from the item analysis included the number of items and number of test takers analyzed for each form, the highest and lowest scores in each form, a measure of internal consistency (coefficient alpha reliability), and a frequency distribution of the number right for each form. Reliabilities were reviewed to confirm that they were consistent with expectations. Typically, reliabilities for routing sections are expected to be about .80 or above because all test takers were administered those items, resulting in wide variability in responses. Lower reliabilities are expected for second-stage forms (for which the restricted variance in overall ability [relative to the whole sample] would be expected to result in lower alpha coefficients) and for forms with relatively few items. Sample sizes were checked for consistency with known values from administrative records from the data collection, item counts were checked for consistency with test specifications, and raw score ranges were also examined.

Most of the assessments used an adaptive two-stage design and therefore required an additional step 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). 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 separately by test form (and separately for the fall and spring of kindergarten), frequency distributions of the total number of items correct (using data from the routing and second-stage forms combined) were examined separately for each form combination (routing + low, routing + middle, and routing + high) 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. (Results of the analysis conducted to determine whether floor or ceiling effects existed in the assessments are presented in chapter 5.)

3.2 Overview of Item Response Theory (IRT)

Measuring the extent of cognitive status and gain, at both the group and individual levels, requires that the assessment forms be calibrated on the same scale within each domain, independently of the particular sample used to obtain those calibrations. IRT procedures were used to carry out such a calibration. There are a number of assumptions that should be examined before applying IRT calibration. Violations of the assumptions of IRT can affect score precision and integrity as well as IRT model fit. First, the sets of test items should be *unidimensional* within a domain with a single, continuous trait (e.g., reading ability) underlying all test form responses. Unidimensionality was studied by a principal components analysis of the assessment items in each domain. Second, the items must demonstrate *local independence*. Local item dependence (LID) can exist when test takers' performances on individual items are correlated once the underlying ability being measured has been controlled for. The local independence assumption is often violated when the answer to a particular question depends (either partially or fully) on knowing the answer to another question, especially when items appear relatively close together in an assessment.

A clear example of local dependence is when a multiple choice question is followed by a constructed response question asking the test taker to explain that child's answer. Such pairs of questions should be scored as a single, combined question. Moreover, if there is information in one item that aids the test taker in answering a different item, those items may demonstrate LID. One consequence of unacknowledged LID is inflated *a* parameter estimates (see below), giving the impression that the item is more discriminating than it really is. LID also may occur in item sets associated with a single prompt such as with passage-based items. LID can be detected using methods such as Yen's Q (Yen 1984) statistic that examine the correlation of item residuals for pairs of items. A third assumption that must be satisfied is that of score *monotonicity*. With monotonicity, the probability of a correct response never decreases as ability increases. Another assumption is that the test is not speeded, meaning that the positions of items relative to the beginning or end of the test does not influence the patterns of response and variability in those items.

Finally, the item function should accurately represent the true relationship between the latent ability being tested and the item responses obtained in the testing. The underlying assumption of IRT is that a test taker's probability of answering an item correctly is a function of that test taker's ability level for the construct being measured and of one or more characteristics of the test item itself. The IRT model enables scoring that uses the pattern of "right" and "wrong" responses to the items administered in a test form, and the difficulty, discrimination power, and probability of guessing each item correctly, to place each test taker at a particular point, θ (theta), on a continuous ability scale.

There are additional requirements when scores from one assessment will be linked to the scores of other assessments, either in the same grade (i.e., fall and spring) or longitudinally. There should be a set of common items shared by different forms or sets of questions, and most, but not necessarily all, content strands should be represented in forms. In a two-stage assessment such as those administered in the ECLS-K:2011, it is also necessary for all children to be administered a common set of items (taking into account both stages) to permit the development of one assessment scale regardless of the second-stage test the child was administered. Additionally, sequential assessments must have increments in difficulty, which can be developed by (a) increasing the problem-solving demands within the same content areas across rounds and (b) 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.

Figure 3-1 is an example of a graph of the logistic IRT function for a hypothetical test item. The graph shows the most general model, the three-parameter IRT model. The three item parameters are *a*

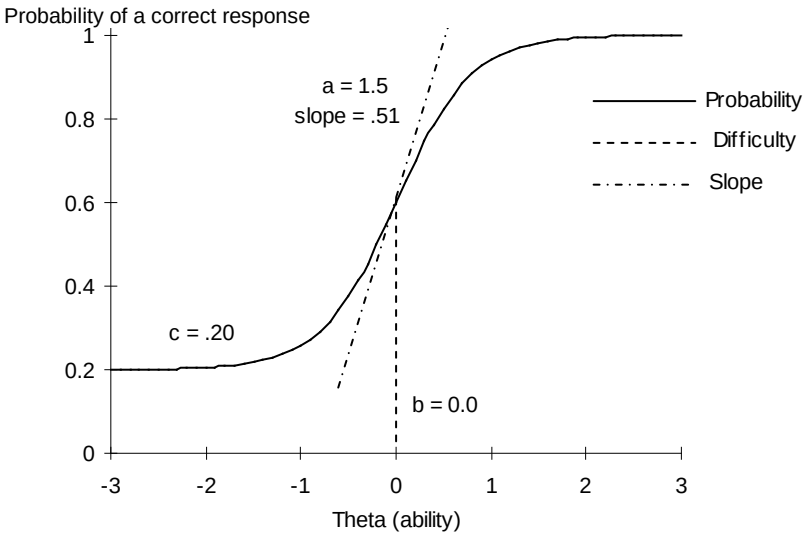
(discrimination), *b* (difficulty), and *c* (pseudo-guessing). The horizontal axis represents the ability scale, theta. The vertical axis represents the probabilities of answering the item correctly given the level of ability (θ). The shape of the curve is given by the following equation, describing 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 = pseudo-guessing associated with item *i*; that is, the probability that a very low-ability test taker will answer item *i* correctly.

Figure 3-1. Three-parameter IRT logistic function for a hypothetical test item



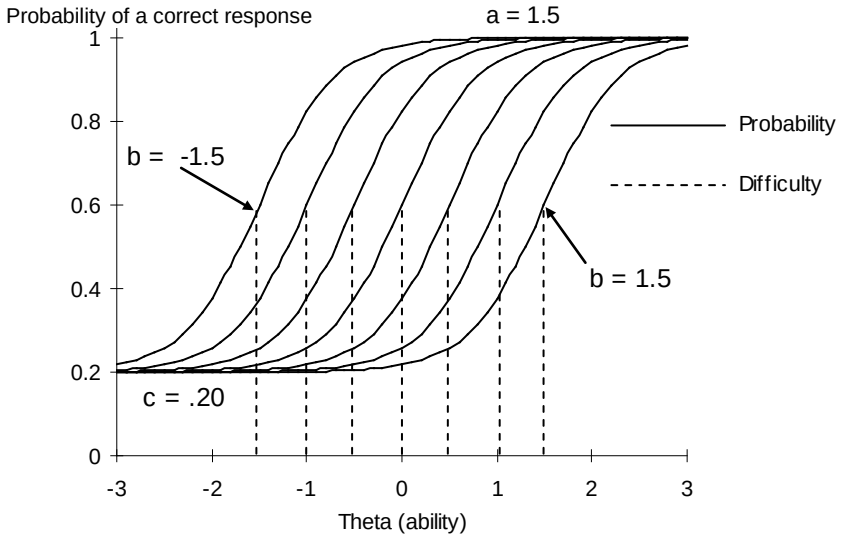
NOTE: *a* = parameter for discrimination; *b* = parameter for difficulty; and *c* = pseudo-guessing parameter.
SOURCE: U.S. Department of Education, National Center for Education Statistics, *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2002-05), 2002.

The IRT *c* parameter represents the probability that a test taker with very low ability will answer a multiple choice item correctly. In figure 3-1, out of 100 people with very low ability, 20 would get the item correct. Note that the *c* parameter does not necessarily equal 1 divided by the number of response options (e.g., 0.25 for an item with four response options). Some incorrect response options may be more attractive than other options (including the correct response), while others may be less likely to be chosen. Therefore, guessing may not be entirely random, and the correct response may not be as likely to be "guessed" as another response option.

The IRT *b* parameter corresponds to the difficulty of the item, which is shown on the horizontal axis in the ability metric, theta (θ). Test takers with ability lower than the item difficulty are less likely to answer the item correctly than test takers with ability higher than the item difficulty. In figure 3-1, this item having difficulty of $b = 0.0$ means that test takers with ability $\theta = 0.0$ have a 60 percent chance of getting the answer correct. In this example, out of 100 people with ability, or theta, equal to 0.0, 60 would be expected to answer the question correctly. The *b* parameter 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 3-2 is an example of a graph of the logistic functions for seven hypothetical 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 items, 60 percent of test takers whose ability level matches the difficulty of the item are likely to answer correctly. The model estimates that 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 $\theta > b$.

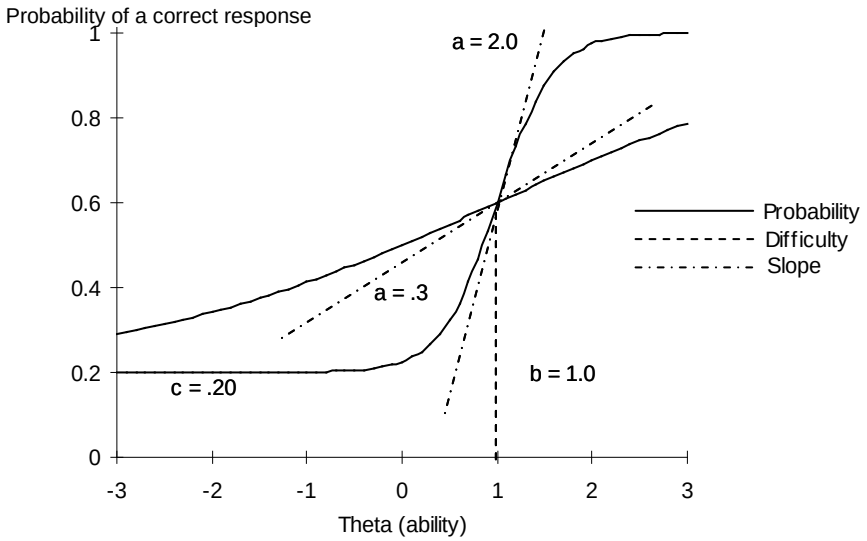
Figure 3-2. Three-parameter IRT logistic functions for seven hypothetical test items with different difficulty (*b*)



NOTE: *a* = parameter for discrimination; *b* = parameter for difficulty; and *c* = pseudo-guessing parameter.
SOURCE: U.S. Department of Education, National Center for Education Statistics, *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2002-05), 2002.

The discrimination parameter, *a*, 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, test takers whose ability level is below the difficulty of the item (i.e., the *b* parameter) from those with ability higher than the item difficulty. By contrast, an item with a relatively flat slope is of less use in determining whether a test taker's correct placement along the continuum of ability is above or below the difficulty of the item. This idea is illustrated by figure 3-3, representing the logistic functions for two hypothetical test items having the same difficulty and guessing parameters but different discrimination. The test item with the steeper slope has a higher discrimination parameter ($a = 2.0$) and, therefore, provides more useful information with respect to whether a particular test taker's ability level is above or below the difficulty level of the item (1.0). In contrast, the flatter curve in figure 3-3 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. Knowing whether a test taker's response to such an item is correct or not contributes relatively little to pinpointing that test taker's correct location on the horizontal ability axis (i.e., that test taker's theta). Thus, a test with highly discriminating items allows for more precise estimation of the test takers' probable ability level than does a test with items that do not discriminate well.

Figure 3-3. 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.
SOURCE: U.S. Department of Education, National Center for Education Statistics,*Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2002-05), 2002.

With respect to evaluating item quality, *a* parameter estimates should ideally be more than 0.5. Items with *a* parameter estimates of 1.0 or above are considered very good. As described earlier, the *a* parameter indicates the usefulness of the item in discriminating between test takers with ability levels above and below the difficulty of the item. The *b* parameter estimates, or item difficulties for the items on a test, 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 test takers. Items that are too easy or too difficult for most of the test takers are of little use in discriminating among them. The *c* parameter estimates (the expectation of a low-ability test taker guessing correctly) tend to be about 0.25 or less for items with four response options, but they may vary with difficulty and, of course, the number of options. Open-ended items typically have a *c* parameter estimate that is close to 0. A two-parameter IRT model, in which the *c* parameter is constrained to be 0, can be used if the likelihood of guessing is very low. In a one-parameter IRT model, items are assumed to discriminate equally well, and the *c* parameter is constrained to be 0. Certain tests can be performed on the data to determine which IRT model (a one-, two-, or three-parameter model) fits the data best.

3.2.1 Calculation of Scores and Treatment of Missing Data

Once there is a pool of test items with parameters that have been estimated on the same scale as the test takers' ability estimates, the probability that a test taker will provide a correct answer for each item in the assessment can be estimated as a function of the test taker's ability estimate, theta, and the estimates of 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 IRT-based score is typically not a whole number.

Before IRT analysis was performed and scores were calculated, a check was run within each domain (reading, Spanish early reading skills [SERS], mathematics, and science) to identify children who had not responded to enough test items to receive a score. Specifically those children who had answered fewer than 10 questions in the assessment for the domain were removed them from analyses.[1] Only items actually attempted by the child were counted toward the scoreability threshold (that is, the minimum of 10 responses needed to calculate a score). A response of "don't know" was treated differently, depending on whether the item was multiple-choice or open-ended. A child's response of "don't know" for a multiple-choice item was treated as an "omit" since the child had the opportunity to select an option but chose not to do so. A response of "don't know" for an open-ended item was treated as an incorrect response since it was interpreted as the child not knowing the answer, and there were no options from which to select. Incorrect responses were treated as valid responses and counted toward the minimum of 10 item responses for scoring. Nonresponses (both omitted and "not reached" items) were not counted. Before being removed from analysis, the data for each child with too few items to score were reviewed visually to verify that too few valid item responses were present. (The counts of children excluded because they had insufficient data are provided in chapter 5.)

3.2.2 Selection of an IRT Model

An issue to be considered when applying IRT methods is the selection of the specific IRT model to be used (i.e., one-, two-, or three-parameter). In general, a one-parameter model has restrictive assumptions that are not easily met, and thus it was not considered for this study's tests. The appropriateness of both the two-parameter IRT model and the three-parameter model was investigated for the ECLS-K:2011 kindergarten assessment data. The analysis included review of outliers, standard errors, thetas, and model fit, and was performed on the data as a whole, and for the items individually. It was

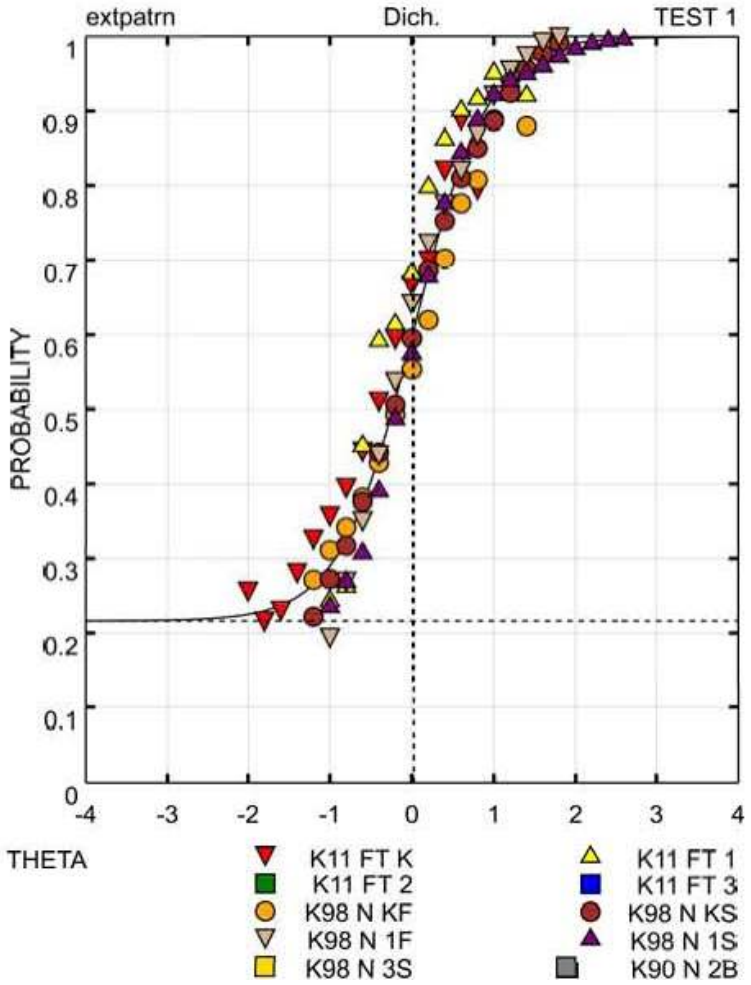
[1] While children who answered fewer than 10 questions technically could have been given a score, when only a few items are available for a child, a stable estimate of child ability is unlikely, leading to a problematic estimate and, possibly, an unreliable estimate of the standard error of measurement.

concluded that the two-parameter and three-parameter IRT models were roughly equivalent in fit. The majority of items for which the fit for the three-parameter model was better than for the two-parameter model were multiple-choice items, where there is a greater likelihood that a child will guess the answer. For the open-ended items, there was a balance between items that were modeled better with the two-parameter model than with the three-parameter model, and vice-versa. Based on the overall review, there was no psychometric advantage to using the two-parameter model, while the three-parameter estimation resulted in a better fit for the multiple-choice items; therefore, the three-parameter model was selected for the ECLS-K:2011.

3.2.3 Evaluating Items Using Empirical Item Characteristic Curves (ICC)

As discussed above, the item parameter estimates can be used to graph the probabilities of correct answers in the ability range that test takers of varying ability will get an item correct. This graph, referred to as an item characteristic curve (ICC), can be used to evaluate how well an item actually performs by adding data points that represent the proportion of correct answers that were given by test takers at all the ability levels represented in the data. This kind of item characteristic curve that includes real data points is called an empirical ICC. The empirical ICC in figure 3-4 shows the fit of the three-parameter model to the actual data for a well-functioning item administered in the assessment field test discussed in chapter 4. Well-functioning items such as this one have data that closely fit the curve and a relatively steep slope at the point of inflection.

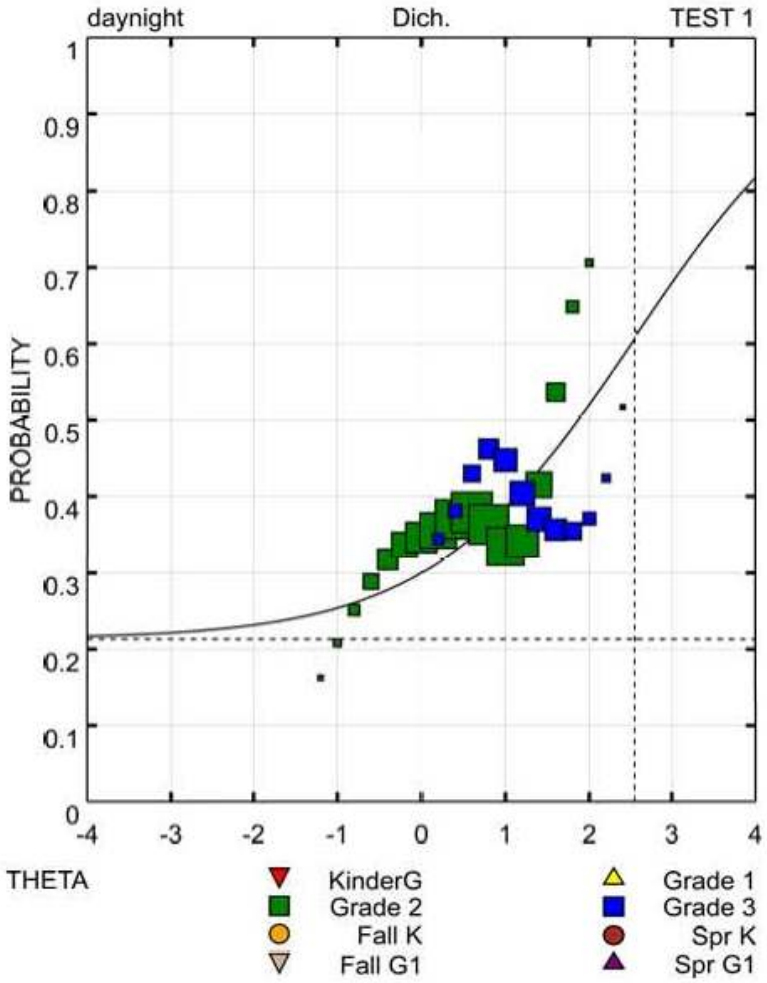
Figure 3-4. Example of an empirical item characteristic curve (ICC) for a well-functioning item: ECLS-K:2011 fall 2009 field test



NOTE: The symbol abbreviations are defined as follows: K11 FT K = ECLS-K:2011 field test, kindergarten sample; K11 FT 1 = ECLS-K:2011 field test, first-grade sample; K11 FT 2 = ECLS-K:2011 field test, second-grade sample; K11 FT 3 = ECLS-K:2011 field test, third-grade sample; K98 N KF = ECLS-K national administration, fall kindergarten; K98 N KS = ECLS-K national administration, spring kindergarten; K98 N 1F = ECLS-K national administration, fall first grade; K98 N 1S = ECLS-K national administration, spring first grade; K98 N 3S = ECLS-K national administration, spring third grade; K98 N 2B = ECLS-K bridge sample, second grade.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

Figure 3-5 shows the empirical ICC of a less successful item included in the assessment field test. Although about 37 percent of the test takers answered this item correctly, performance on this item is not strongly related to overall ability: throughout most of the ability range, test takers were about equally likely to answer correctly, so it does not discriminate well. This item also violates the monotonicity assumption, because higher ability test takers appear to be less likely to answer correctly than lower ability test takers at certain points on the ability scale.

Figure 3-5. Example of an empirical item characteristic curve (ICC) for a poorly functioning item: ECLS-K:2011 fall 2009 field test



NOTE: The symbol abbreviations are defined as follows: KinderG = ECLS-K:2011 field test, kindergarten sample; Grade 1 = ECLS-K:2011 field test, first-grade sample; Grade 2 = ECLS-K:2011 field test, second-grade sample; Grade 3 = ECLS-K:2011 field test, third-grade sample; Fall K = ECLS-K national administration, fall kindergarten; Spr K = ECLS-K national administration, spring kindergarten; Fall G1 = ECLS-K national administration, fall first grade; Spr G1 = ECLS-K national administration, spring first grade.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

3.2.4 Item Information and Measurement Precision

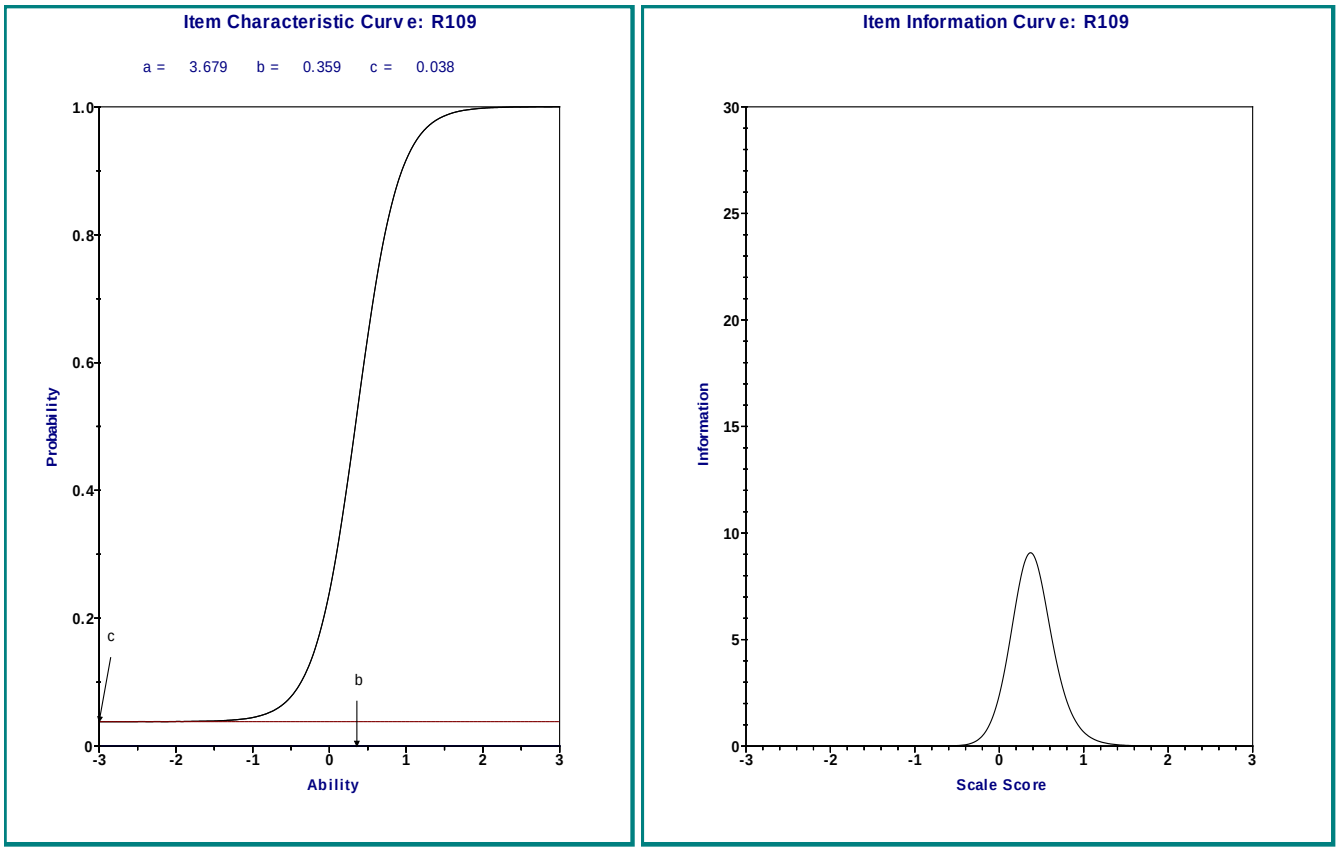
Another way to measure item and test quality is to estimate the item information function (IIF) and test information function (TIF). In psychometrics, the precision of parameter estimates at the various ability levels can be measured using the information function. This is computed as a function of the reciprocal of the measurement error, denoted as σ^2. Equation 3.2 defines the information function (I):

$$I = \frac{1}{\sigma^2}. \tag{3.2}$$

When evaluating test data using IRT, estimating the ability parameter, or θ, of each test taker is of primary interest. If the test contains a large number of highly discriminating items with a difficulties spread across the range of test takers' scores, each test taker's true ability can be estimated 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 or too easy for a particular test taker, a precise estimate of that test taker's θ, or ability level, cannot be obtained. In this situation, the variance of estimates (measurement error) will be relatively high, and the value of the information function will be relatively low. Therefore, the information function tells how well each ability level is being estimated. It is computed for each item answered by a test taker.

Much as the ICC provides a visual representation of item functioning in terms of the estimated *a*, *b*, and *c* parameters, the IIF provides a visual representation of the place on the ability scale where the item measures best. Figure 3-6 shows the ICC and IIF for a hypothetical item. This item has good discrimination and seems to measure well for test takers with a theta ability of approximately 0 to 1.

Figure 3-6. Item characteristic curve (ICC) compared to item information function (IIF)



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

The definition of the item information function depends on the IRT model used. For the three-parameter model, the item information function I_i is defined as

$$I_i(\theta) = a^2 \frac{Q_i(\theta)\left[P_i(\theta) - c\right]^2}{P_i(\theta)(1-c)^2} \tag{3.3}$$

where

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

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

θ = ability of the test taker;

$P_i(\theta)$ = probability that a test taker of ability θ will answer the item correctly; and

$Q_i(\theta)$ = 1.0 minus the probability that a test taker of ability θ will answer the item correctly.

The aggregate of all the individual IIFs is the test information function and thus will be much greater than any single item information function. Therefore, it would be expected that a test measures ability more precisely than does a single item. The test information function is estimated using only the administered items with valid responses. Generally, the more items answered, the greater the precision in estimating ability. In addition, more information is derived from items with high discrimination, or *a* parameter estimates; therefore, for a test with a range of items with high *a* parameter estimates across the appropriate range of difficulty levels, the test information function will show high levels of information across the child ability range.

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. The equation for the test information function is

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

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.

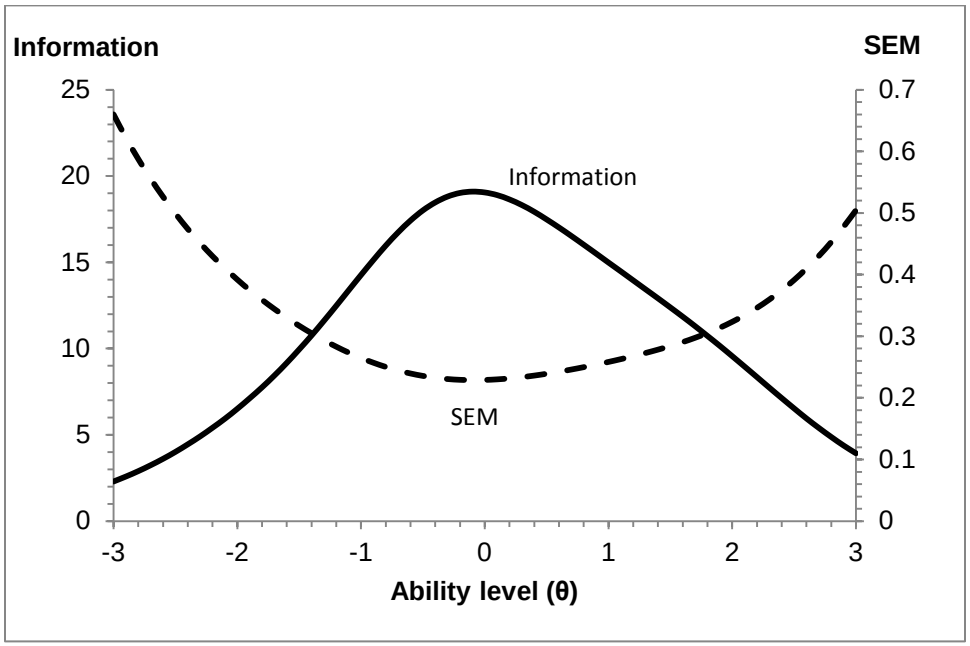
Tests are designed with item difficulties that are matched to the expected ability levels of the target population of test takers. There are generally more middle-difficulty items matching the ability of the majority of test takers, and relatively few easy and difficult items designed for the test takers 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.

Since the overall test is used to estimate the ability level of the child, the test information function is used to estimate the standard error, which is often referred to as the standard error of measurement, or *SEM*. The standard error is estimated from the reciprocal of the square root of the test information function:

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{3.5}$$

An example TIF is shown in figure 3-7. Overall, this hypothetical test seems to measure well through the -2 and +2 theta ability range. The solid line in this graph represents the information, while the dashed line is the reciprocal of the square root of that information, the standard error (*SEM*). The *SEM* is conditional on ability; as the information increases, the standard error decreases.

Figure 3-7. Example test information function (TIF)



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

The *SEM* was estimated for each theta estimate for each child for each domain. These estimated standard errors are provided in the data files for each of the thetas.

3.2.5 Item Response Theory Estimation Using PARSCALE

The PARSCALE (Muraki and Bock 1991) computer program for estimating IRT models was used for estimating item parameters and estimating test takers' ability levels on a scale that was then used to produce scale scores based on the whole item pool. This section provides a general description of the PARSCALE program. Appendix B includes more detail about the preparation of scored-item files for use in PARSCALE and how PARSCALE estimates the IRT model.

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. Expectation-maximization steps are performed until the largest change in item threshold or slope parameters is less than the convergence criterion value

(0.005), or the maximum number of cycles has been reached (300). The convergence criterion and maximum number of cycles are set using values defined by ETS standard practices. Comparison of the IRT-estimated probability of a correct response with the actual proportion of correct answers to a test item for test takers 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.

In the ECLS-K:2011, as well as other longitudinal growth studies, multiple subpopulations of the same group of children are defined by abilities measured at differing times.[2] That is, after the fall and spring assessments had been completed, there were two recognizable subpopulations of different ability levels related to time of testing (i.e., data collection round). The fall subpopulation will have, on average, a lower expected level of performance than that found for the same children during the spring data collection. At the time of the fall assessments, only a few children may be able to answer the more challenging questions, such as those related to reading passages. Thus, there may be a limited set of data collected on the most difficult items in each domain. However, many of these more difficult items were re-administered to some of the study children and administered to others for the first time in the spring assessment. As a result, when the two data rounds were combined for IRT analyses, the spring data were used to stabilize the parameter estimates from fall, especially for the more difficult items in the item pool.

A strength of the PARSCALE and other approaches to IRT is that they can incorporate prior information about the ability distribution (i.e., data from the previous round) in the current round ability estimates. This is particularly crucial for measuring change in longitudinal studies. Pooling all available information—that is, pooling all item responses for all test takers at both within-grade time points and recalibrating all of the item parameters using Bayesian priors[3] reflecting the ability distributions associated with each particular round—provides for an empirically based adjustment of estimated item parameters and ability scores to values more representative of the population than the data from one round taken in isolation might suggest (Muraki and Bock 1991). Bayesian priors (also typically referred to simply as "priors") are essentially a priori distributional assumptions about proficiency and have relatively little influence on the estimation of proficiency if there is sufficient information collected from a test taker; they have more influence if the test taker's information is sparse.

[2]As used here, "subpopulation" refers to the data available at a point in time or around a given ability level (e.g., fall or spring kindergarten). As used in IRT, subpopulation divides all available data across data rounds (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.

[3]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.

Using the total item pool in conjunction with the selected Bayesian priors (which reflect the ability distributions associated within each grade-level round) leads to a reduction in extreme values for the item parameter estimates, resulting in a reduced likelihood of perfect and chance scores based on the scoring methodology used. This, in turn, makes it more likely that gains can be measured even in the upper and lower tails of the distribution. Each round of data collection (fall and spring) is treated as a separate subpopulation with an independent 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 round compared with another, the extremes of the ability estimate in the round 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 rounds) 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 within-grade longitudinal scales are determined by the items, and the initial Bayesian prior ability means for the children in the different rounds are in turn determined by the differential performance of the children on these items across rounds. The approach of using adaptive testing procedures combined with Bayesian procedures that allow for the use of prior values on both ability distributions and the item parameter estimates is needed in longitudinal studies to minimize ceiling and floor effects.

3.3 Construct Validity: Assessing Dimensionality

An essential requirement in the applicability of IRT is that a test is unidimensional, meaning that the items included in the test all contribute to measuring a single underlying construct. For example, the kindergarten science assessment is designed to measure a unitary science knowledge and skills and does not provide adequate detail on distinct constructs in science at that age level, such as classification

skills versus observation skills. To investigate the dimensionality of the cognitive assessments, principal components analyses were run on the data collected during national data collection. In each domain, principal components of the item correlation matrix were computed to check for the presence of a single dominant component, as well as the percentage of variance explained by the first and each subsequent components. Rotations were carried out for two to five components. Component loadings, which are correlation coefficients for each item with each hypothesized component, were then examined by specialists to determine whether high loadings on any but the first component suggested that the test might be multidimensional.

Ideally, to define unidimensionality, the ratio of the first component to the second component should be at least 3:1.[4] If the ratio of components does not establish the single-component status unequivocally, the next step is to look at the component loadings and examine the content of the items that load on different components. If the items cluster according to difficulty and not content (i.e., the easiest items generally load on one component, and harder items load on different components), this would suggest that, although the content may vary (e.g., vocabulary, conventions of print, letter and sound skills, sight words, and comprehension), the differences in performance are likely due to a strong underlying single component. With true multiple components suggesting multidimensionality, sets of items along a fairly wide range of difficulty and content would be clustering on different components.

3.4 Group Differences in Item Functioning

Assessment items showing an unexpectedly large difference in item performance between subgroups when the two groups are matched on their total score (e.g., Black and White children with a score of 50 on the reading assessment) should be examined for bias and excluded from scoring if it is determined that differential performance on the item is *unfairly* associated with subgroup membership (that is, if the difference exists because of an attribute of the item not related to the construct being measured). For example, in the case of a mathematics item administered in both the English and Spanish versions of the assessment, if differential performance was shown by children matched on total score, it might be determined that the translation resulted in a favoring or disfavoring of a language group.

The ECLS-K:2011 assessment data were examined for bias using several procedures that assessed differential item functioning, or DIF. First, items were evaluated for statistical DIF, or purely

[4] Based on ETS standard practices.

empirical evidence of differential item functioning. Two statistical DIF methods were used in detecting differential performance of subgroups on the ECLS-K:2011 direct cognitive assessments during each round. 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-K:2011 direct child cognitive assessments, specifically the fact that not all children received the same items or items of the same difficulty, made number-right scores inappropriate for use as stratifying, or blocking, variables. Instead, the IRT ability estimate, theta, was used as the stratifying variable, divided into 41 equally spaced intervals.[5] 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 whether the child got a given item correct or incorrect. 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 responding correctly to the item in question. If the likelihood is not the same, it is possible that the item functions differently for reasons other than ability, and the item should be reviewed further to determine whether it was biased. Although the M-H statistic looks at the correct response rates for two groups while controlling for total score, no assumptions need to 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 responding correctly to 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

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

because of this range. A more common measure of difficulty is the proportion-correct or *p* value. Test developers 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. A classification scheme that uses the M-H Delta Difference, or M-H D-DIF, as an effect size for DIF was used in the analyses of the ECLS-K:2011 data. 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 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.

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. The proportion correct metric is defined as the comparison of the proportions correct for 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.

In the P-DIF procedure, the proportion correct for each group is calculated at each score level. P-DIF uses a weighting factor at each score level to weight differences in the proportion correct between the focal group and the reference group. The use of this same set of weights for both groups is the essence of the standardization approach. 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, and the M-H D-DIF uses more complex statistics in detecting DIF. The two procedures yield measures that are highly correlated (typically .9 and above); if discrepancies are

observed, they are typically found for very easy and very 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.

However, any 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. As Cole and Moss (1989) point out, items demonstrating statistical DIF 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 attend schools where the curriculum emphasizes learning letter names and sounds, while others attend schools where relatively more time is spent reading stories to the children. Both groups may have similar total scores in reading, but the letter recognition items may be significantly more difficult for one group than for the other. Therefore, the fact that an item is identified by these DIF procedures as functioning differently does not mean that the item is necessarily unfair to any particular group. DIF procedures are merely statistical screening steps that indicate that the item is behaving somewhat differently for one or more subgroups.

The second step in examining assessment data for bias 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 statistical DIF, either in favor of the reference group or against the reference 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 retained in scoring. If the reviewers find otherwise, the item is removed from the scoring procedures.

DIF procedures were carried out after each round of the ECLS-K:2011 assessments. Items were checked for differential functioning using the child's sex and race/ethnicity, and, after the spring kindergarten round, round of administration 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: non-Hispanic Black children, Hispanic children of any race, and non-Hispanic Asian children (including Native Hawaiians and Pacific Islanders). There were too few non-Hispanic 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. 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 are discussed in detail in chapter 5.

3.5 Evaluating Common Item Functioning During the Development of the Kindergarten Longitudinal Scale

The study of the relationships between children's early childhood experiences at kindergarten entry and their gains in academic skills during the kindergarten year required the development of a vertical assessment scale spanning kindergarten that had optimal measurement properties throughout the achievement range. That is, the assessments administered in the fall and spring together needed to reflect the core curriculum elements covered across the entire kindergarten year, and scores from each round needed to be compared to one another. It was possible to meet these two requirements by ensuring that the test forms used in fall and spring had common items, and that there was overlap in the difficulty distributions of the items for each round. The common items tie the vertical scale together across rounds. As a general rule, for the average size of these assessments, at least 20 percent of

the items should overlap between adjacent rounds. The same assessment was used in the fall and the spring, so this criterion was met.

However, although the content and presentation of each of the common items were identical in the fall and spring, it is still possible for the items to function differently. Since common items exist on adjacent second-stage test forms, some children are administered the same item on two different forms in subsequent rounds, in a different item order and among a different set of items. Of course, it is expected that performance on the items would improve as children acquire new skills and knowledge, and thus an increase in the probability of a child giving a correct answer for any given item would be observed. However, as the children gain skills, their ability increases so they have a higher probability of answering the items correctly than they may have had on the same items in a previous year. The difficulty of items in the context of the entire assessment for a given domain should be maintained for the common items used to anchor the scale.

To assess the common functioning of overlapping items in each domain, the data from the fall and spring kindergarten data collections were pooled and preliminary estimates of IRT item and ability parameters were obtained, using all items from the assessment forms from each round. Each common item was initially assumed to be common functioning, and this assumption was tested using differential item functioning procedures described in the previous section. The round of administration contrast group was defined as fall (reference group) compared with spring (focal group). Items that are not common functioning would not be used as common items for the purposes of developing a vertical scale. These items would be treated as completely different items in the calibration and scoring by round, unlike common items, which are treated as the same item administered in each round of data collection Results of the DIF analyses are included in chapter 5.

3.5.1 Concurrent Calibration and Computation of Final Scores

The sections in this chapter outline the processes used to convert the raw data collected from children to assessment scores that are provided on the ECLS-K:2011 data file. After the data were cleaned, preliminary item parameters were estimated using IRT procedures. These preliminary estimates were used in analyses examining item bias and common functioning. Once these analyses were completed, final parameter estimates were calculated for the set of items retained within each domain.

Within grade, each of the rounds of data collection—fall and spring—is treated concurrently as a separate subpopulation with its own ability distribution for the purpose of IRT calibration. As described above, all item responses from each round of data collection are pooled into a single calibration, with data from each round retaining a separate ability distribution. This treatment, which is a feature of PARSCALE and other approaches to IRT, when using a Bayesian option, provides for an empirically based shrinkage toward subpopulation means for extreme ability estimates, both low and high. This shrinkage, which was discussed earlier in section 3.2.5, 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 are derived from the IRT item parameter estimates and ability estimates. As described in section 3.2 and illustrated in equation 3.1, the set of three parameter estimates for each item defines a logistic function corresponding to the estimated probabilities of correct answers for test takers across the ability scale. These estimated probabilities are summed over all items to get a scale score representing an estimate of the number of items the test taker would be expected to answer correctly in each domain. At each time point, the ability estimates are used in combination with the item parameter estimates to generate an estimated probability of a correct response for each item, summed over all unique items in each domain, for each round. For example, a test taker who is tested at both kindergarten rounds (fall and spring) will have two ability estimates and the associated scores for each round.

This page intentionally left blank.

4. DEVELOPMENT OF THE TWO-STAGE COGNITIVE ASSESSMENT TEST FORMS

This chapter provides information about the development of the direct child cognitive assessments in reading, mathematics, and science for the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), from the initial item pool development to the finalization of the assessment forms used in the national kindergarten data collections. The executive function assessments are not discussed in this chapter; the study administered existing, well-tested assessments so no development or field testing for items in this domain were needed.

4.1 Development of the Item Pool

During the design phase for the national assessments, a pool of assessment items tapping knowledge and skills in the content categories outlined in chapter 2 was developed, evaluated, and field tested. Items were developed and assembled to span kindergarten entry through the spring of third grade.

In the first step of this process, the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) and the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) cognitive assessment batteries were reviewed to identify items that were appropriate for a new cohort of kindergartners and the current education environment. The assessment developers looked for items that measured grade-appropriate knowledge and skills according to current state and national curriculum standards. Items that were deemed appropriate were brought forward for inclusion in the ECLS-K:2011. These items allow comparisons to be made between two cohorts of kindergarten students who entered school more than a decade apart. Items that were deemed no longer relevant were dropped or revised. It was also necessary to develop some new items because the existing ECLS-K and ECLS-B batteries lacked items covering topics within the standards and new areas of interest to the research community that were identified through discussions with experts on the Technical Review Panel (TRP), and reviews of recent research. Items in these new areas of interest will help the study provide data relevant to emerging areas of interest. For example, in reading, items assessing phonological awareness, including segmentation, substitution, and blending, and those assessing rhyming were item types that were included in the ECLS-K:2011 but not the ECLS-K.

Additionally, new assessments measuring English basic reading skills (EBRS) for all children, and Spanish-speaking children's reading skills[1] and knowledge in their native language (the Spanish early reading skills, SERS assessment), were developed for the ECLS-K:2011. Once the initial pool of items was developed for the reading assessment in English, reviewers selected those tapping more basic knowledge and skills for potential inclusion in the EBRS. Items were then translated into Spanish for field testing and potential inclusion in the SERS.

After the initial item pool was identified, all of the items and assessments proposed for the ECLS-K:2011 were reviewed by measurement specialists and content area specialists (i.e., reading assessment specialists reviewed all the reading items proposed for each grade's assessment, mathematics assessment specialists reviewed the mathematics items, science assessment specialists reviewed the science items, and specialists in early reading skills and bilingual assessments reviewed the EBRS and SERS items). The content experts were given information about the study background, the objectives of the national assessments (described in chapter 2), and a description of the adaptive two-stage design that would be used in the study. They also received copies of the proposed items and guidelines for their review. The guidelines asked the reviewers to evaluate the quality of each item in the initial item pool using the following criteria:

- Estimated difficulties for the items should be appropriate for the targeted grade levels, and the group of items selected for each grade level should include a portion of items that are somewhat easier, and others that are somewhat more difficult, to permit measurement in the extremes of the ability distribution.
- The items should be accurate: the content should be correct, the presentation (including the language, illustrations, and charts) should be exact, and the spelling and grammar should be precise.
- Multiple-choice questions should have a single unambiguously right answer, and all others should be unambiguously wrong.
- Open-ended questions should be fairly easy for an assessor to score as correct or incorrect.
- Incorrect options should be plausible responses to the question (i.e., options that could plausibly be selected by a test taker who does not know the answer). Ideally, a test taker who knows the material should get it correct, and one who does not should only be able to guess at random and not be able to eliminate answers that are obviously impossible.

[1] This measure assessed Spanish reading skills and knowledge for Spanish-speaking children who were not sufficiently proficient in English to be assessed in English.

- There should be nothing about the phrasing or the context of the item that is tricky or confusing (e.g., use of units in a question that is not designed to measure familiarity with the particular units may be problematic for some students and interfere with their being able to answer the question).

- Avoid using items that may possibly be easier or more difficult for a particular subgroup of children in a way that is unrelated to their overall ability.

- The question and response formats should not give hints about what the correct answer is (e.g., in a measurement item, the units in each response option should match those in the question to avoid selection of a response option based on units only).

- The content of the items should be both appropriate for the grade level being assessed (as defined by the framework) and important to measure, per the recommendations of the content specialists.

- The item content should be characteristic of the typical curriculum at the intended grade level.

- The presentation of the items (context, language, illustrations, response options) should be optimal for measuring children's ability in each domain.

In addition to these general considerations, there were specific issues to consider for each cognitive domain. For example, in the reading domain, the reviewers considered if grade-appropriate language in the passage text was used, if the text was relevant for children and could sustain a child's interest at the targeted grade level, and if the sight word items were considered high frequency (i.e., recognizable and used regularly) for the targeted grades. In mathematics, for example, reviewers were asked to evaluate the possible impact of language load (that is, the complexity of phrases and sentences used in the item); it was anticipated that the items would be translated to Spanish, so items with significant language load that could not be minimized were excluded from the field test. The impact of language load is not only relevant for the Spanish-translated version of the mathematics assessment, but for the English version as well, where the confounding influence of language ability may result in concerns about validity: that is, whether the test is actually measuring mathematics ability and not reading ability. Similarly, reviewers of the science items were asked to identify items that may measure reading and/or mathematics ability more so than science ability (e.g., an item requiring a child to make a calculation from experimental results may be a measure of both science and mathematics abilities). Reviewers of the SERS items were asked, for example, to identify items that would likely translate well from English to Spanish, and to select items with limited language load to aid in the translation.

The items also underwent a sensitivity review by reviewers trained to detect objectionable material. They reviewed the items to ensure that they did not inadvertently reflect gender or ethnic

stereotypes or inappropriate assumptions about people with disabilities; have a lack of racial, ethnic, or gender diversity among the characters in stories or test items; or have content that may be deemed offensive or inappropriate with respect to any such groups or individuals.

Based on the comments of the expert reviewers, some items were revised. The final pool of items to be field tested was selected from the subset of those items approved or revised by the expert reviewers and that had the best psychometric properties. More details about the item selection criteria can be found in section 4.2.

4.2 Field Test Design

In the fall of 2009, two field tests were conducted to test the assessment items being considered for inclusion in the direct child assessments for the kindergarten, first-, and second-grade collections of the national study. These field tests served as the primary vehicle for estimating the psychometric properties of items in the assessment battery item pool and producing psychometrically sound and valid direct cognitive assessment instruments. Each field test focused on different components of the assessment. The primary goal of the English field test, which focused on the assessments administered in English, was to collect data (specifically, item statistics) to inform the development of the kindergarten, first-, and second-grade assessments for reading, mathematics, and science. A secondary goal was to collect child rating data from teachers for the development of an academic rating scale indirectly assessing children's science skills. The primary goal of the Spanish field test was to estimate the psychometric parameters of each of the EBRS and SERS items for Spanish-speaking children and establish whether or not these items could be used to produce valid measures for both an *English* reading score and an assessment of early reading skills (e.g., letter recognition and sounds) *in Spanish* for these children. Although the data in the Spanish field test were used to evaluate how the EBRS items performed for Spanish-speaking children, the performance of these same items for non-Spanish-language speaking minority children was not determined, as non-Spanish-language speaking minority children were not included in the field test sample. Details about the field test samples and methodology can be found in the field test report included in appendix A.

4.2.1 English Field Test Design

Assessment items in reading, mathematics, and science were administered in the fall 2009 field test to 2,978 children (905 kindergartners, 846 first-graders, 818 second-graders, and 409 third-graders). Although the purpose of the field test was to test items for the assessments from kindergarten through second grade, a limited sample of third-graders was included in order to provide item measurement information for high-ability second graders. The 279 items in reading (including 28 passages with 134 associated items), 146 items in mathematics, and 171 items in science were distributed across multiple forms targeted for administration to either kindergartners and first-graders or second- and third-graders. Two forms in mathematics, two forms in science, and four forms in reading were assembled for each grade combination (kindergarten/first grade and second/third grade). The forms, which were designed to be approximately parallel within grade with respect to the content and difficulty of the items, were then sorted into 8 booklets per grade combination, totaling 16 booklets in all. As shown in table 4-1, each booklet contained one reading form and one form for either mathematics or science. The forms were spiraled, or ordered, such that each of the subject forms appeared as the first section of the assessment in some booklets and as the second section of the assessment in other booklets. The sixteen booklets were administered to approximately equal numbers of children within each grade grouping. The booklet spiraling and field test sample size resulted in approximately 300–800 observations for each test item, with the number of observations (or children who received the item) dependent upon the distributions of the items across forms within and across grades. On average, reading and mathematics items had about 500 observations, and science items had about 400 observations. Items appropriate for administration to children in any grade included in the field test were presented in both kindergarten/first grade and second/third grade booklets, which resulted in more observations for those items. Items that appeared in fewer forms had fewer observations. Both multiple-choice and open-ended items were presented in each form for each subject area.

Table 4-1. Organization of booklets: ECLS-K:2011 fall 2009 field test

Booklet	Observations	Section 1	Section 2
K/1st Grade Booklet 1	221	K/1st Reading 1	K/1st Mathematics 1
K/1st Grade Booklet 2	223	K/1st Reading 2	K/1st Mathematics 2
K/1st Grade Booklet 3	223	K/1st Reading 3	K/1st Science 1
K/1st Grade Booklet 4	216	K/1st Reading 4	K/1st Science 2
K/1st Grade Booklet 5	217	K/1st Mathematics 1	K/1st Reading 1
K/1st Grade Booklet 6	217	K/1st Mathematics 2	K/1st Reading 2
K/1st Grade Booklet 7	218	K/1st Science 1	K/1st Reading 3
K/1st Grade Booklet 8	212	K/1st Science 2	K/1st Reading 4
2nd/3rd Grade Booklet 1	156	2nd/3rd Reading 1	2nd/3rd Mathematics 1
2nd/3rd Grade Booklet 2	152	2nd/3rd Reading 2	2nd/3rd Mathematics 2
2nd/3rd Grade Booklet 3	155	2nd/3rd Reading 3	2nd/3rd Science 1
2nd/3rd Grade Booklet 4	156	2nd/3rd Reading 4	2nd/3rd Science 2
2nd/3rd Grade Booklet 5	153	2nd/3rd Mathematics 1	2nd/3rd Reading 1
2nd/3rd Grade Booklet 6	151	2nd/3rd Mathematics 2	2nd/3rd Reading 2
2nd/3rd Grade Booklet 7	146	2nd/3rd Science 1	2nd/3rd Reading 3
2nd/3rd Grade Booklet 8	147	2nd/3rd Science 2	2nd/3rd Reading 4

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), field test, fall 2009.

Data collected from all children in the field test were analyzed together, regardless of grade level, since the emphasis was on evaluating the performance of the items across a broad range of ability levels and maintaining maximum sample sizes to help stabilize estimates. When evaluating item characteristics for a specific grade's assessment (e.g., the assessment for the kindergarten rounds), such as the difficulty of the items, the focus was predominantly, but not exclusively, on the data collected from children in the particular grade in question.

4.2.1.1 Reading Field Test Forms and Items

Each of the four kindergarten/first-grade reading field test forms had 65 items and five to seven short reading passages about which some of the questions were asked. Similarly, the second-/third-grade reading field test forms had 70 items and nine or ten short reading passages. Several passages and associated items were presented on multiple forms within and across grades. This overlap of items across test forms was designed to stabilize item parameter estimates and provide a strong link between the forms to better enable the selection of test items that would successfully measure gain in successive rounds. Some of the passages and items were newly developed for the ECLS-K:2011 field test; others were taken from the ECLS-K kindergarten/first-grade and third-grade assessments. Items that had been part of the

ECLS-K reading assessments were included in the ECLS-K:2011 field test forms as a cost-savings measure (since fewer new items needed to be developed) and to support linking across the studies.

4.2.1.2 Mathematics Field Test Forms and Items

The field test contained 146 unique mathematics items, divided among two forms each for the kindergarten/first-grade and second-/third-grade combinations and designed to be approximately parallel within grade with respect to the content and difficulty of the items. Each form appeared in two different test booklets as either the first or second assessment along with a reading form. Some items appeared in multiple forms within or across grades. Some items that had been part of the ECLS-K mathematics assessments were included in the ECLS-K:2011 field test forms to support linking across the studies.

4.2.1.3 Science Field Test Forms and Items

The field test contained 55 unique science items, divided among two forms each for the kindergarten/first-grade and second-/third-grade combinations and designed to be approximately parallel within grade with respect to the content and difficulty of the items. Similar to mathematics, each science form appeared in one test booklet paired with a reading form, as either the first or second assessment. Some science items appeared in multiple forms within or across grades. As with reading and mathematics, some items were also previously administered in the ECLS-K assessments. However, in the ECLS-K a general knowledge assessment including both science and social studies items was administered in kindergarten and first grade. Thus, the number of science items that could be used was relatively limited, so it was necessary to develop relatively more new science items for these grade levels compared to reading and mathematics.

4.2.2 Spanish Field Test Design

The fall 2009 Spanish field test served as the primary vehicle for (a) estimating the psychometric parameters of all items in the English and the Spanish early reading skills assessments and (b) producing psychometrically sound and valid direct assessment instruments of English basic reading

skills and Spanish early reading skills. The psychometric parameter estimates of the English items were evaluated from data collected from Spanish-speaking children to check for common functionality with the same items administered in English to English children, while the psychometric parameter estimates of the Spanish early reading skills items were evaluated to check functioning for only Spanish-speaking test takers. The goal was to assess 1,200 Spanish-speaking kindergartners regardless of their English proficiency. Approximately 50 schools in five geographic areas were selected for the Spanish field test. The geographic areas were selected based on the percentage of enrolled students in the school who were Hispanic, according to the Common Core of Data Public Elementary/Secondary School Universe Survey: School Year 2006–07. Assessment items measuring early reading skills were administered in the Spanish field test. The EBRS section, consisting of 28 items in English, was followed by the SERS section, consisting of 46 items in Spanish. Both sections were administered (along with two simple practice items in English) to the 1,115 kindergarten children participating in the Spanish field test.

4.2.2.1 English Basic Reading Skills (EBRS) and Spanish Early Reading Skills (SERS) Field Test Items

The 28 EBRS items administered in the field test included items administered previously in the ECLS-K and ECLS-B assessments, as well as items newly developed for the ECLS-K:2011. The items assessed letter and sound knowledge, phonological awareness, and vocabulary. These EBRS items were also administered in the English field test so that information about how these items perform for both English- and Spanish-speakers could be obtained. The 28 EBRS items were translated into Spanish and included in the SERS. In addition, the SERS included more complex items measuring knowledge of sight words and print conventions administered in Spanish. In total, the SERS included 46 items. Similar to the item pools included in the English field test, the item pools tested in the Spanish field test included both multiple-choice and open-ended questions.

EBRS items were evaluated as part of the full reading assessment item evaluation for the English field test sample because these items are a portion of the full reading assessment. EBRS items administered in English were also analyzed in a separate calibration with the EBRS items administered in Spanish from the Spanish field test. A separate calibration for the EBRS items administered in English and Spanish was required since the original assessment plan included utilizing the EBRS items to route children into the English or Spanish assessment, as appropriate. However, the analysis results showed that there was no specific cut score that could be developed using the EBRS items only to route children into

the English or Spanish versions of the assessment. (Subtests of the Preschool Language Assessment Scale (*pre*LAS) were used for that purpose.)

4.3 Field Test Results and the Development of the National Assessments for the Kindergarten Data Collection

Data collected during the field tests were used to evaluate item quality and identify flaws in wording or response options, ascertain the range of ability likely to be encountered in the sample of students who would take the national assessment, and calibrate the field test item difficulties on the same scale as student achievement, so that items of appropriate difficulty could be selected for the final forms. In addition, the performance of both English and Spanish items included in the Spanish field test was evaluated. Data collected in the English and Spanish field tests were evaluated separately.

The remainder of this chapter discusses the adequacy of the field test item pool and the design of the national assessments for the kindergarten data collection, including how items were selected for inclusion in the national reading, SERS, mathematics, and science assessments. The discussion of the analysis references psychometric methods and uses psychometric terms that are explained in chapter 3.

4.3.1 Methods Used to Analyze the Field Test Data

As noted above, the items field tested in the fall of 2009 were drawn from several sources: the ECLS-K kindergarten and first-grade assessment,[2] the ECLS-K third-grade assessment,[3] and the ECLS-B preschool and kindergarten assessments.[4] Each of these sources had a large number of items available for use in the ECLS-K:2011 national assessments. In addition, 118 reading, 80 mathematics, and 142 science items were newly developed for the ECLS-K:2011 to measure concepts not included in the earlier studies, a subset of which were used in the field test forms. By design, the majority of items field tested had been used before, either in the ECLS-B or the ECLS-K, so concerns about item quality had already been largely addressed for these items. In both the English and Spanish field test analyses, attention was paid to data from the items newly developed for the ECLS-K:2011 and how the items that

[2] Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, and spring 2000.
[3] Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.
[4] Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Preschool and Kindergarten National Assessments, fall 2005 through spring 2007.

were previously developed performed nearly a decade later. In the Spanish field test analysis, additional analyses were performed examining the effect of the language of administration on item function.

In order to measure each child's status accurately in the national assessment, it is important that each child receives a set of test items that is appropriate for that child's skill level. The selection of items for the national administrations involved consideration of two sets of estimates: the difficulty parameters for each of the items in the pool, and the range of children's ability levels that was expected in each round. 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 item response theory (IRT) analysis. IRT calibration of the English field test item data was carried out for each subject area by pooling data from the following sources of data:

- ECLS-K:2011 2009 field test, kindergartners (approximately 890 cases);
- ECLS-K:2011 2009 field test, first graders (approximately 850 cases);
- ECLS-K:2011 2009 field test, second graders (approximately 800 cases);
- ECLS-K:2011 2009 field test, third graders (approximately 400 cases);
- ECLS-K fall kindergarten national data collection (approximately 18,000 cases);
- ECLS-K spring kindergarten national data collection (approximately 19,000 cases);
- ECLS-K fall first-grade national data collection (approximately 5,000 cases);
- ECLS-K spring first-grade national data collection (approximately 16,000 cases);
- ECLS-K spring second-grade bridge sample[5] (approximately 900 cases); and
- ECLS-K spring third-grade national data collection (approximately 14,000 cases).

The ECLS-K did not have a separate science assessment in the kindergarten and first-grade rounds; therefore, for the calibration and evaluation of the science items, the science items from the K–1 general knowledge assessment were selected and pooled with science items used in later grades of the ECLS-K and new science items developed for the ECLS-K:2011. Although selected items from the ECLS-B were included in the ECLS-K:2011/2009 field test assessments, the ECLS-B sample was not

[5] Due to budgetary constraints, data were not collected in second grade in the ECLS-K study. However, a bridge sample of second graders was assessed to establish a longitudinal scale between the first and third grade ECLS-K national assessments. More details on the bridge sample may be found in *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–62).

included in the calibration because its data collection spanned an entire year and the purpose of this analysis was to define expected item performance and child ability levels at two distinct time periods, fall and spring.

A separate IRT calibration focusing on the effects of language of administration at the kindergarten level was conducted for the Spanish field test data. For this calibration, data collected from kindergartners in the ECLS-K:2011 field test (approximately 1,000 cases) were pooled with data from the kindergarten and first-grade rounds of the ECLS-K. Data from later rounds of the ECLS-K were not used in analyses of the Spanish field test data because the EBRS and SERS items measured more basic knowledge skills most appropriate for inclusion in a kindergarten assessment, and it was not anticipated that the Spanish-translated items would be administered beyond the spring of first grade.

Spanish-translated mathematics items were not included in the Spanish field test administration. In the ECLS-K, the Spanish-translated mathematics items were shown to perform similarly to the English versions of the items; thus, field testing of the ECLS-K:2011 Spanish-translated mathematics items was deemed unnecessary.

Pooling of the data for IRT calibrations was done for two primary reasons. First, for analyses of data from both the English and Spanish field tests, the items included in two or more of the datasets mentioned above serve as anchors, so that parameter estimates for items and the mean ability levels of the test takers measured using the different assessments could all be put on a common scale. Data from the ECLS-K:2011 field test provide a link between the newly developed field test items and items from the older ECLS-K assessments, enabling items from all cohorts to be put on a common scale necessary to make direct comparisons. Second, the IRT model used requires at least 400 examinees per item in order to obtain stable item parameter estimates. As noted earlier, the relatively small sample of the fall 2009 field test resulted in some items having fewer observations than this minimum needed for IRT analyses. Pooling the field test data with the large samples from the ECLS-K data collections also serves to stabilize parameter estimates that would lack precision if the data from only the field test were used to evaluate the assessment items. Even with pooling, some new items, especially those of high difficulty, were more frequently omitted by the child or not administered because they appeared at the end of the assessment form and were not reached before the end of the assessment period,[6] resulting in fewer observations than ideal. In these cases, the items were not included in the national administration

[6] For the field test, the assessment time for each child was limited to approximately one hour. Assessors administered as many items as possible within that hour.

calibrations, as the item parameter estimates were less robust than those with higher numbers of observations.

Although the datasets are pooled, the samples are identified individually so that the ability range of each sample can be obtained separately. The mean and standard deviation of the ability levels for each of the samples were calculated based on data from the pooled sample. Therefore, an estimated ability range for the target administrations (e.g., fall and spring kindergarten, 2010–11) can be determined.

The pool of items available for assembly of the national test forms was not limited to the items in the 2009 field tests. Using the methodology described, the difficulty parameters for all of the items used in all of the datasets were estimated on a common metric, regardless of whether the items were tested in the 2009 field test. Thus, virtually *all* items in the source tests were considered part of the item pool for the purpose of test assembly for the ECLS-K:2011 national data collection.

4.3.2 Criteria Guiding the Selection of Items for the National Kindergarten Assessments

The item selection process was guided by numerous objectives, including the following:

- **Psychometric characteristics:** Selecting items that discriminate well across the full range of ability levels and do not show differential item functioning (DIF).
- **Difficulty:** Matching the difficulty of the test questions to the expected range of ability that would be found in the national administrations; choosing items for the routing and second-stage forms that were of appropriate difficulty; avoiding floor and ceiling effects.
- **Test specifications:** Within each subject area, matching the percentages of items pertaining to each content category in the assessment to the target percentages specified in the assessment framework as closely as possible. A primary goal of item selection for the national assessment was to include items from each content category in the same proportions as indicated in the framework specifications. However, the ability to meet this goal depended on the number of available items in each category that had good psychometric characteristics and fell within the identified difficulty ranges.
- **Horizontal linking and vertical scaling:** Having a sufficient number of items that are administered to all children in the router and that are shared among second-stage forms within a data collection round so that one stable scale can be established for measuring status in that round, having a sufficient number of items that are shared among assessments across rounds so that one stable scale can be established for

measuring gain across rounds, and having an adequate number of items that had been fielded in the ECLS-K to permit cross-cohort comparisons.

- **Assessor feedback:** Incorporating recommendations made by the field staff based on their observations of how children responded to the items and the ease or difficulty of the assessment administration.

- **Time limits:** Making efficient use of testing time, both to limit cost and to minimize burden on test takers and schools.

4.3.2.1 Estimated Ability Levels for the ECLS-K:2011 Kindergarten National Sample and Target Ranges for Item Difficulties

The kindergarten assessments were designed to support measurement of children's skills and abilities in reading, mathematics, and science as accurately as possible, at all levels of ability found within the ECLS-K:2011 fall and spring kindergarten rounds, and to include items that would also be appropriate for use in subsequent rounds of data collection to support vertical scaling. IRT ability estimates were used to define targeted difficulty ranges for the different forms of each subject area assessment in kindergarten. The ability (theta) estimates for the ECLS-K:2011 fall kindergarten field test sample and the ECLS-K fall and spring kindergarten national samples were estimated from the pooled data described earlier and were used to estimate the range of children's abilities that could be expected in the ECLS-K:2011 national fall and spring kindergarten data collections.

The analysis using the pooled data showed differences in the estimated mean ability levels between the ECLS-K:2011 field test and the ECLS-K national samples. The mean ability level for ECLS-K national sample in the fall of kindergarten was about 0.6–0.7 of a standard deviation (based on either the ECLS-K:2011 field test or ECLS-K standard deviation) below the mean ability level of the fall 2009 kindergarten field test sample for reading and mathematics, and about 0.2–0.3 of a standard deviation below for science (using only the science items from the ECLS-K general knowledge assessment for the ECLS-K sample). The differences in mean ability level for each subject area between the ECLS-K national fall first-grade sample and the ECLS-K:2011 field test first-grade sample were similar.

Several factors may contribute to these differences observed for mean ability level. First, the ECLS-K:2011 field test schools were selected to include a diverse group of schools and students, but they

were not selected in a way that would ensure that the sample was perfectly representative of the population. It is possible that the ECLS-K:2011 field test sample included a disproportionate number of children with higher than average ability, resulting in the range of abilities in the field test sample being somewhat attenuated at the low end compared with the full population. Another possible factor could be real changes in the kindergarten population in the interval between 1998 and 2009 with respect to prior exposure to early learning experiences. Expansion of preschool programs serving disadvantaged children could result in kindergartners in 2009 being, on average, better prepared for school than those who entered kindergarten in 1998. Without knowing the explanation for the discrepancy with certainty, the range of difficulty of the test forms was targeted to be suitable for a range of ability levels defined by the low end of the ECLS-K distribution as the lowest ability level and the upper end of the ECLS-K:2011 field test distribution as the highest ability level. This range, from roughly two standard deviations below the fall ECLS-K mean ability level to two standard deviations above the estimated spring ECLS-K:2011 mean ability level,[7] was expected to include at least 95 percent of children in the ECLS-K:2011 national sample, even if the difference between the ECLS-K:2011 field test and the ECLS-K national results remained unexplained. Another reason for extending the difficulty range of the items at both the low and high end of the ability range was to avoid floor and ceiling effects in the national assessment.

The anticipated range defines not only the ability range of the children, but also the corresponding difficulty parameter estimates of the items required for the assessment. The estimated range of theta was used to define the range of abilities targeted by the national test forms. Thus, the process of choosing test items relied on matching the difficulty of the items to the abilities of the test takers. To optimize the measurement accuracy of the tests, the selected items were approximately equally spaced along the ability/difficulty scale. Items that fell outside the targeted ability/difficulty range for kindergarten generally were not considered for inclusion in the national kindergarten assessments except when needed to avoid floor and ceiling effects, or to provide additional overlap between forms to support development of a common score scale.

In addition to the full range of difficulty for the entire assessment, separate ranges of difficulty had to be estimated for low-, middle-, and high-ability groups in reading and mathematics so that items could be selected for the routers and the three second-stage tests. For each estimated ability

[7] The spring kindergarten ECLS-K:2011 mean theta was calculated by assuming the growth from fall kindergarten to spring kindergarten to fall first grade would be the same as what was observed in the ECLS-K. That is, the children in the ECLS-K sample gained a little more than one standard deviation from fall to spring kindergarten, and nearly another half standard deviation by fall first grade. This is consistent with what was observed in the ECLS-K:2011 field test sample, which showed slightly less than 1.5 standard deviations difference between the average fall kindergarten theta and the average fall first-grade theta.

range, the low end of the range was computed using the mean ability level and the associated standard deviation of the lower scoring sample (ECLS-K national data collection), while the high end of the range was based on the mean ability level and the associated standard deviation of the ECLS-K:2011 field test sample. Generally, the lowest ability level ranged from two standard deviations below the ECLS-K national mean to the ECLS-K:2011 field test mean; the middle ability level ranged from one standard deviation below the ECLS-K national mean to one standard deviation above the ECLS-K:2011 field test mean; and the highest ability level ranged from the ECLS-K national mean to two standard deviations above the ECLS-K:2011 field test mean. The router was designed to have items with difficulties spanning the entire expected range of ability, because having information about child's performance on items with different difficulties was necessary to determine to which second-stage test the child should be routed. Items with difficulty in the ranges noted above were selected for each second-stage test.

By design, the ranges of ability overlap for two main reasons. First, the overlap in the ability range covered by each form results in an overlap in items selected for the second-stage tests. As noted above and in chapter 3, such overlap is necessary to develop one stable scale for the entire assessment. Second, it assures that reliable scores can be calculated for instances in which a child is routed to a second-stage test that is not exactly matched to his or her level true level of ability. For example, a child whose true ability falls within the defined range for the lowest level second-stage form could be routed to the middle-level second-stage form because he guessed correctly on one router item, resulting in the lowest total router score that directs children to the middle form. Having lower-level items in the middle form allows for the estimation of that child's ability even though the majority of the items he received in the middle form might have been too difficult for him. Conversely, a child whose true ability falls within the defined range for the highest-level second-stage form could be routed to the middle-level second-stage form because he was tired and not paying close attention to questions he could have answered correctly but did not. Having higher-level items in the middle form allows for the estimation of that child's ability even though the majority of the items he received in the middle form might have been relatively easy for him.

In each domain, items with acceptable psychometric properties that fell within the targeted ability/difficulty range were sorted into sets according to content category and similarity of presentation. In reading, for example, "basic skills" items were presented in several different formats (e.g., some letter sound items were administered by asking the child to point to which letter made the sound produced by the administrator, while others asked the child to produce the sound associated with a letter.) The difficulty statistics were reviewed for the items within each content/presentation type set, and each set

was identified as being suitable for the single-stage form in science and SERS, or the routing, low, middle, and/or high form in reading and mathematics, according to the difficulty of the majority of the items in each set. Different presentations of the same content were compared, and where there was redundancy, the item sets with the strongest psychometric characteristics were selected. In general, the similar content/presentation types were ordered in increasing order of average difficulty (although most had a spread of difficulty within types). Other factors were also taken into consideration, such as grouping items by format to minimize changes in task instructions.

While the parameter estimates from the field test were provisional and were recalibrated after the national administration, the estimates produced by the calibration of the ECLS-K:2011 field test and ECLS-K national data allow for relatively accurate preliminary estimates of the item parameters for the purposes of test-form development. There are many factors that may contribute to slight differences between the parameter estimates calculated from the field test data and those calculated from the national administration data, for example: the assortment of items selected for the national test; the number and location of practice items; the use of discontinued rules in the national data collection but not in the field test; and real differences between the field test sample, which was not necessarily nationally representative, and the population of kindergarten students in 2010 and historical samples.

4.3.2.2 Item Quality and Reliability

To contribute useful information about children's skill level, test items selected for the final forms should ideally have high *r*-biserials (0.3 or higher) and IRT discrimination (*a*) parameter estimates (1.0 or higher), as well as a good fit of the IRT model to the empirical data. Items with high discrimination parameter estimates permit accurate placement of estimates of theta on the ability continuum. A few of the selected items fell short of these standards but were selected for the national assessments for other reasons such as coherence with framework specifications, overlap with the ECLS-K assessments, or links to a selected reading passage. 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.

4.3.3 Composition of the Final National Kindergarten Assessments

4.3.3.1 Reading

Overall, the field test items for the reading assessment performed well. The item analysis showed that the majority of items had *r*-biserials that were well above the desired value of 0.3. Of the 279 unique items administered, 31 exhibited *r*-biserials lower than ideal. Of those 31 items, 12 were very easy items (P+ values greater than 0.9), and another 12 were quite difficult (P+ values less than chance for the multiple-choice items, or less than 0.1 for open-ended items).[8] The remaining 7 items exhibited low *r*-biserials because children who chose the correct response had an average score that was just slightly higher than those who chose an incorrect response, suggesting that the items may not have had a single, clearly correct response. These 7 items of moderate difficulty that had low *r*-biserials were not selected for inclusion in the national assessment. Of the 24 items that were very easy or very difficult, 6 were included in the national assessment in order to prevent floor or ceiling effects.

The remaining 248 items showed the expected trends in response selection; that is, the correct response was more likely to be selected by students who had higher average scores than by students who had lower average scores.

Review of the IRT plots showed good fit of item data with the estimated parameters. Although the fit was good for most of the items, the discrimination was not necessarily so. This pattern of good fit with poor discrimination generally occurred with items that were either relatively easy or hard. In selecting items for the national assessment, items with poor fit or discrimination were avoided.

4.3.3.1.1 English Basic Reading Skills (EBRS)

The EBRS items were evaluated as part of the full reading assessment item evaluation for the English field test sample because these items are a portion of the full reading assessment. These items served as part of the reading routing test, with additional routing items[9] administered to children who performed well on the EBRS set. The items on the EBRS were relatively easy and were not adequate to distinguish between children at middle- and high-ability levels, and thus a second, more difficult set of

[8] For items that are either very easy or very difficult for most of the sample, correct or incorrect responses are not highly correlated with ability level, resulting in low *r*-biserials.
[9] Also referred to as Part 2 Routing.

routing items was needed. The EBRS items, as well as the other reading items, are discussed in the sections below.

As discussed earlier, the items included in the EBRS also were evaluated separately using data from both the English and Spanish field tests in order to estimate the psychometric parameters of the items and to check for common functionality with the same items administered in English to English-speaking children with those administered in Spanish to Spanish-speaking children. This separate calibration was required, as stated earlier, to assess the viability of using the EBRS items to route children into the English or Spanish assessment, as appropriate. It was shown the items were not adequate for this purpose.

The majority of field test items for the EBRS performed well in the Spanish field test (using the same criteria described above). The item analysis showed that all items except one exhibited *r*-biserials well above the desired value of 0.3. The one item with an *r*-biserial slightly below ideal was quite difficult with a P+ value near 0.1. This item was not included in the national assessment. Review of the IRT plots showed high discrimination and good fit of item data with the estimated parameters.

4.3.3.1.2 Item Difficulty

Table 4-2 lists the estimated means and standard deviations of ability level (theta), all calibrated on the same scale, for kindergartners from the different data collections that were used to calculate the full range of ability levels (and, therefore item difficulties) that needed to be covered by the assessment. Table 4-3 shows the estimated ability ranges for the entire assessment as well as for the low-, middle-, and high-ability level groups for fall and spring kindergarten. Table 4-3 also shows the number of items selected for the national assessment that have a difficulty falling within the peak range (within two standard deviations of the mean) for each second-stage form. Note that not all items fall within the peak range in the second-stage forms. Items outside the peak range are intentionally included to extend difficulties beyond the peak range and to provide additional overlap between forms to support development of a common score scale.

Table 4-2. Estimated means and standard deviations of reading ability level (theta) for children in kindergarten

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Fall kindergarten – ECLS-K:2011 field test	-0.60	0.63
Fall kindergarten – ECLS-K national data collection	-1.03	0.62
Spring kindergarten – ECLS-K:2011 field test (estimated)	+0.03	0.63
Spring kindergarten – ECLS-K national data collection	-0.39	0.59

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2009 field test, and Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998 and spring 1999 national data collections.

Table 4-3. Peak difficulty ranges for the national kindergarten reading assessment, routing plus second stage: ECLS-K:2011

	Fall kindergarten			Spring kindergarten		
Item	Anticipated low-level abilities (-2*SD* to mean)	Anticipated mid-level abilities (-1*SD* to +1*SD*)	Anticipated high-level abilities (mean to +2*SD*)	Anticipated low-level abilities (-2*SD* to mean)	Anticipated mid-level abilities (-1*SD* to +1*SD*)	Anticipated high-level abilities (mean to +2*SD*)
Estimated ability range	-2.2 7 to -0.60	-1.65 to +0.03	-1.03 to +0.66	-1.57 to +0.03	-0.98 to +0.66	-0.39 to +1.29
Number of items with difficulties in anticipated peak ability range:						
Part 1 Routing (EBRS)	13	15	10	15	9	2
Part 2 Routing	0	8	20	8	20	20
Form A	11	9	2	7	2	0
Form B	8	9	8	7	8	5
Form C	1	1	11	1	11	19

NOTE: *SD* = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

4.3.3.1.3 Adherence to Framework Specifications

As described in section 2.1.1, the content categories included in the framework specifications for the ECLS-K:2011 reading assessment are the following:

- basic skills;
- vocabulary;
- locate/recall;
- integrate/interpret; and
- critique/evaluate.

Design of the reading assessment is somewhat different from the other domains since the items associated with reading passages are selected in sets rather than individually. Only a limited number of passages could be included in any assessment form, because the time for the assessment was relatively limited and the child needed to read the passages before answering the questions. For efficiency, when selecting items, the test developers tried to include as many questions associated with each reading passage as possible. A reading passage was favored for inclusion in the national assessment if it had one or more associated items in one of the more difficult content categories, such as **integrate/interpret** or **critique/evaluate.** However, some of the passages also had associated items in the **basic skills** category, which were already overrepresented because of the administration of the EBRS assessment. Thus, the need to include several items associated with a given reading passage affected the distribution of items across content categories on the test as a whole.

Table 4-4 provides information about how the final reading assessment developed for the national kindergarten rounds of data collection compares to the framework specifications in terms of the distribution of items by content category. The table indicates the targeted percentage within each content category, as well as the actual percentage and number of items selected for the national administrations within each content category. The need to include the full set of 20 EBRS items to measure basic reading skills in English for all children, regardless of English language proficiency, skewed the content category percentages to include more basic skills items than would otherwise be necessary. For this reason, table 4-4 shows a targeted percentage of items by content category compared with the actual percentages both with and without the required EBRS items.

Table 4-4. Framework targets and items by content area for the national kindergarten reading assessment: ECLS-K:2011

Content area	Targeted percent of items	Including EBRS items		Excluding EBRS items	
		Actual number	Actual percent	Actual number	Actual percent
Total	100	83	100	63	100
Basic skills	50	53	64	35	56
Vocabulary	15	11	13	9	14
Locate/recall	20	14	17	14	22
Integrate/interpret	10	3	4	3	5
Critique/evaluate	5	2	2	2	3

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

The passage sets chosen for the high difficulty second-stage form were selected to maximize the number of i**ntegrate/interpret** and **critique/evaluate** items of appropriate difficulty for the fall and spring kindergarten assessments. However, as can be seen in the information presented in table 4-4, even with this maximization, the percentage of items in each of these categories fell short of the targets. The available item pools did not include enough items in these categories that performed well for children who did not yet have the skills required to read and understand text. The percentage of items in the **locate/recall** category fell close to the targeted percentage, whereas the percentage of items from the **vocabulary** category was slightly lower than targeted due to the limitations of the item pool.

The percentage of items from the **basic skills** category was slightly higher than targeted, and even more so when the EBRS items are included in the counts. The ECLS-K:2011 incorporates items that cover a wider variety of skills within the **basic skills** content category than the original ECLS-K. Table 4-5 lists the different subcategories of **basic skills** items in the kindergarten national forms and the numbers and percentages of items in these subcategories. (The numbers in table 4-5 are based on the items in the kindergarten national forms *with* the EBRS items included.) Items in the phonemic awareness, segmentation, blending, and rhyming subcategories (a total of 12 items) were not included in the original ECLS-K assessments. Although the framework specified that items measuring children's knowledge of syllables be included, items asking the child to indicate the number of syllables in a word did not perform well in the field test. They were too difficult, did not have good psychometric properties

and, according to assessor feedback, were confusing to kindergartners. Thus, the kindergarten national assessment did not include any items about syllables.

Table 4-5. Subcategories of basic skills items included in the national kindergarten reading assessment: ECLS-K:2011

Basic skills subcategory	Number of items	Percent of items
Total	53	64
Letter sounds	4	4.8
Beginning sounds	5	6.0
Ending sounds	4	4.8
Phonemic awareness	6	7.2
Phonemic substitution	2	2.4
Segmentation	2	2.4
Blending	2	2.4
Rhyming	2	2.4
Sight words	12	14.5
Print convention	8	9.6

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

As stated above, the inclusion of passage item sets in the assessment, the addition of the phonemic item subcategory to the **basic skills** category, and the requirement that all children be administered the EBRS item set resulted in the actual distribution of the items across content categories deviating from the targets in the assessment framework specifications.

4.3.3.1.4 Spanish Early Reading Skills (SERS)

Overall, the field test items for the SERS performed well. The item analysis showed that the majority of items had *r*-biserials that were well above the desired value of 0.3. One difficult vocabulary item and three quite difficult reading items at the end of the assessment exhibited *r*-biserials lower than ideal. As with the EBRS, review of the IRT plots showed good discrimination and fit of item data with the estimated parameters.

The ability estimates for the ECLS-K:2011 Spanish field test sample, estimated from the pooled analysis of the Spanish field test data described earlier in section 4.3, were used to estimate the

range of abilities in Spanish early reading that could be expected for the Spanish-speaking English language learner (ELL) children in the ECLS-K:2011 national sample in kindergarten. Assuming the SERS field test sample was reasonably representative of the national sample, it was expected that the range from roughly two standard deviations below the mean to two standard deviations above the mean would include about 95 percent of the ECLS-K:2011 national sample.

As was done for the other domains, some items with *b* parameter estimates above two standard deviations on the high end were included to avoid ceiling effects. However, of the items that functioned well for the SERS, the one with the lowest *b* parameter estimate was just about two standard deviations below the mean ability level. As a result, it was not possible to include items with difficulty more than two standard deviations below the mean to try and prevent floor effects, although it was not anticipated that there would be many children with an ability level below the difficulty level of the easiest item. Table 4-6 shows the number of items included in the SERS with a difficulty level within specific difficulty ranges defined incrementally by standard deviation. The five items at the upper end of the estimated ability/difficulty range (*b* > +2 SD) were included to avoid a ceiling effect and also to address the possibility of significant growth from fall to spring kindergarten in Spanish early reading skills.

Table 4-6. Number of items in the national kindergarten Spanish early reading skills (SERS) assessment, by difficulty range: ECLS-K:2011

Difficulty range (*b*)	Number of items
Entire assessment	31
b < = -2 *SD*	0
-2 *SD* < b <= -1 *SD*	6
-1 *SD* < b <= mean	8
mean < b <= +1 *SD*	7
+1 *SD* < b <= +2 *SD*	5
+2 *SD* < b	5

NOTE: SD = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

The framework design for the SERS consists entirely of individual **basic skills** and **vocabulary** items. Table 4-7 provides information about how the final SERS assessment developed for the national kindergarten rounds of data collection compares to the framework specifications in terms of the distribution of items for the two content categories included. The targeted percentages for each of

these categories in the SERS reflect the relative proportion of these two categories in the English reading assessment. The target percentages were met exactly.

Table 4-7. Framework targets and items by content area for the national kindergarten Spanish early reading skills (SERS) assessment: ECLS-K:2011

Content area	Targeted percent of items	Actual number of items	Actual percent of items
Total	100	31	100
Basic skills	77	24	77
Vocabulary	23	7	23

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

The majority of items in the SERS assessment are from the **basic skills** category. Table 4-8 lists the different subcategories of **basic skills** items in the form, and the numbers and percentages of items that pertain to each subcategory. The phonemic awareness items (substitution, segmentation, blending, etc.), as well as the ending sounds items, were excluded from the SERS due to concerns that they would not be testing the same skills when translated, given the differences in English and Spanish sounds. Though there was no requirement that the targeted percentages for each content category for the SERS match the targets for the English reading assessment exactly, the percentages from the English assessment were used to guide the selection of items for the SERS. As a result, the percentages of items that fall in each content subcategory in the SERS and the English reading assessment are similar.

Table 4-8. Subcategories of basic skills items included in the national kindergarten Spanish early reading skills (SERS) assessment: ECLS-K:2011

Basic skills subcategory	Number of items	Percent of items
Total	24	100
Letter recognition	5	21
Letter sounds	3	13
Beginning sounds	2	8
Sight words	7	29
Print convention	7	29

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

4.3.3.2 Mathematics

As with reading, the field test items for mathematics performed well. The item analysis showed that the majority of items had *r*-biserials that were above the desired value of 0.3. Of the 146 unique mathematics items administered in the field test assessment, 10 had *r*-biserials lower than ideal, either because the item was very easy (N = 6) or very difficult (N = 4).

4.3.3.2.1 Item Difficulty

Table 4-9 lists the estimated means and standard deviations of ability level (theta), all calibrated on the same scale, for kindergartners from the different data collections that were used to calculate the full range of ability levels (and, therefore item difficulties) that needed to be covered by the mathematics assessment. Table 4-10 shows the estimated ability ranges for the entire assessment as well as for the low-, middle-, and high-ability level groups for fall and spring kindergarten. Table 4-10 also shows the number of items selected for the national assessment that have a difficulty falling within the peak range (within two standard deviations of the mean) for each second-stage form. As with the design of the reading forms, the range of difficulty for the selected items was extended at both the low and high ends to avoid floor and ceiling effects. For this reason, some items with difficulty parameter estimates below -2.25 at the low end and above +1.16 at the high end were included in the national mathematics assessment.

Table 4-9. Estimated means and standard deviations of mathematics ability level (theta) for children in kindergarten

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Fall kindergarten – ECLS-K:2011 field test	-0.60	0.53
Fall kindergarten – ECLS-K national data collection	-0.99	0.63
Spring kindergarten – ECLS-K:2011 field test (estimated)	+0.10	0.53
Spring kindergarten – ECLS-K national data collection	-0.40	0.59

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009, and Kindergarten Class of 1998-99 (ECLS-K), fall 1998 and spring 1999.

Table 4-10. Peak difficulty ranges for the national kindergarten mathematics assessment, routing plus second stage: ECLS-K:2011

Item	Fall kindergarten: Anticipated low-level abilities (-2*SD* to mean)	Fall kindergarten: Anticipated mid-level abilities (-1*SD* to +1*SD*)	Fall kindergarten: Anticipated high-level abilities (mean to +2SD)	Spring kindergarten: Anticipated low-level abilities (-2*SD* to mean)	Spring kindergarten: Anticipated mid-level abilities (-1*SD* to +1*SD*)	Spring kindergarten: Anticipated high-level abilities (mean to +2*SD*)
Estimated ability range	-2.25 to -0.60	-1.62 to -0.07	-0.99 to +0.47	-1.58 to +0.09	-0.99 to +0.62	+0.40 to +1.16
Number of items with difficulties in anticipated peak ability range:						
Routing	10	12	13	13	13	7
Form A	14	9	3	8	3	0
Form B	14	19	17	18	18	9
Form C	2	5	10	7	14	23

NOTE: SD = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

4.3.3.2.2 Adherence to Framework Specifications

As discussed in section 2.1.2, the content categories included in the framework specifications for the ECLS-K:2011 mathematics assessment are the following:

- number sense, properties, and operations;
- measurement;
- geometry and spatial sense;
- data analysis, statistics, and probability; and
- patterns, algebra, and functions.

Table 4-11 provides information about how the final mathematics assessment developed for the national kindergarten rounds of data collection compared to the framework specifications in terms of the distribution of items by content category. The actual percentages of items matched the targeted percentages in the **geometry and spatial sense** and **data analysis, statistics, and probability** content areas. The shortfall in the **measurement** category is due to the lack of **measurement** items in the item pool that fell within the difficulty range suitable for kindergarten and had good psychometric characteristics. Conversely, the mathematics assessment included more items from the **patterns, algebra, and functions** and **number sense, properties, and operations** categories than targeted. The distribution of item difficulties required that these additional **patterns, algebra, and functions** and **number sense properties, and operations** items be included to ensure accurate assessment across the ability distribution.

Table 4-11. Framework targets and items by content area for the national kindergarten mathematics assessment: ECLS-K:2011

Content area	Targeted percent of items	Actual number of items	Actual percent of items
Total	100	75	100
Number sense, property and operations	75	57	76
Measurement	5	2	3
Geometry and spatial sense	3	2	3
Data analysis, statistics, and probability	8	6	8
Patterns, algebra, and functions	9	8	11

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

4.3.3.3 Science

Compared with the reading and mathematics items, fewer of the field test items for science performed well. Of the 171 unique science items, 72 exhibited *r*-biserials lower than the desired value of 0.3. The majority of items administered to first- and second-grade students functioned well, and thus provided an adequate pool from which to develop assessments for these grade levels. However, many of the field-tested items were not appropriate to administer in a kindergarten assessment, because they were too difficult (especially for children in the fall of their kindergarten year) or they were not common functioning across grade levels (i.e., for kindergartners and first-graders), or both. (Items that are

common-functioning across rounds permit longitudinal measurement of gains, as these common-functioning items are used as anchors.)

These field test findings indicated that science knowledge and skills could not be validly and reliably assessed in the fall of kindergarten with the items that had been developed and tested. Although some items did function well when administered to kindergartners in the fall during the field test, there were not enough items to develop a full assessment for the fall. Moreover, the items that did have acceptable performance were predominantly from a single content category (life science), making it impossible to select a set of items for a two-stage kindergarten science assessment that would have been consistent with the test framework. Instead, a limited, 20-item, single-stage test was developed for administration in the spring kindergarten data collection so that some science assessment data would be collected during the base year. The assessment included relatively easy items that functioned well for kindergartners or first-graders, or both. The items that functioned well for first-graders were included under the assumption that if they functioned well in the field test for first-graders in the early fall of their first-grade year, they would be appropriate for administration to kindergartners in the spring of their kindergarten year during the national data collection.

Review of the IRT plots showed good fit of item data with the estimated parameters for the majority of items, but the discrimination was not necessarily so. In selecting items for the national assessment, items with poor fit or discrimination were avoided.

4.3.3.3.1 Item Difficulty

Table 4-12 lists the estimated means and standard deviations of ability level (theta), all calibrated on the same scale, for kindergartners from the different data collections that were used to calculate the full range of ability levels (and, therefore item difficulties) that needed to be covered by the science assessment. As with the reading and mathematics assessments, items with difficulty parameter estimates below the anticipated lowest theta (-1.70) and above the anticipated highest theta (1.35) were included to avoid floor and ceiling effects.

Table 4-12. Estimated means and standard deviations of science ability level (theta) for children in the spring of kindergarten

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring kindergarten – ECLS-K:2011 field test (estimated)	+0.05	0.65
Spring kindergarten – ECLS-K national data collection	-0.04	0.83

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), field test, fall 2009, and Kindergarten Class of 1998-99 (ECLS-K), spring 1999.

4.3.3.3.2 Adherence to Framework Specifications

As described in section 2.1.3, the content categories included in the framework specifications for the ECLS-K:2011 science assessment are the following:

- Scientific inquiry;
- Physical science;
- Life science; and
- Earth science.

Table 4-13 provides information about how the final science assessment developed for the national spring kindergarten data collection compares to the framework specifications in terms of the distribution of items by content category. The actual percentages of items match the targets exactly in all categories.

Table 4-13. Framework targets and items by content area for the national kindergarten science assessment: ECLS-K:2011

Content area	Targeted percent of items	Actual number of items	Actual percent of items
Total	100	20	100
Scientific inquiry	25	5	25
Physical science	25	5	25
Life science	25	5	25
Earth science	25	5	25

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

4.3.4 Performance Simulations and Cut Scores Used for Routing

Once the items were selected and allocated to the routing and low-, middle-, and high-level forms for the national assessment, simulations of performance on the routing test and second-stage tests were run in order to calculate cut scores for the routing test that would determine which second-stage form children would be administered. To conduct the simulations, 10,000 thetas (ability estimates) were randomly drawn from a normal distribution with a mean and standard deviation corresponding to the expected fall and spring kindergarten ability levels in each domain. For each randomly generated theta, the probability of a correct response was computed for each item on the routing and low-, middle-, and high-level forms, separately for each subject.

Next, an estimated number-right score was determined for each theta by summing the probabilities of a correct response for the items on each test form. This procedure never results in a score of zero because for the multiple-choice items, the probability of a correct response is always greater than zero, due to guessing. To address this limitation on the score calculation, a random number between 0 and 1 was also generated for each item. This was done so that an integer number-right score could be computed for use in the estimation of cut scores and in review of floor and ceiling effects. If the random number generated was less than or equal to the predicted probability of a correct response, the item was scored correct (=1); the item was scored incorrect (=0) if the random number was greater than the predicted probability of a correct response. For example, if the probability of a correct response estimated from the item parameters and an individual theta was .9 and the random number generated was .5, the

item would be scored correct. This is a logical procedure because if the probability of correctly answering an item is .90, in most administrations the item would be scored correct. Conversely, if the probability of a correct response was .1 and the random number generated was .5, the item would be scored incorrect. Again, since the probability of correctly answering an item is only 10 percent, in most administrations the item would be scored incorrect. Summing the zeros and ones from these calculations resulted in integer scores for each form for each subject. Cross-tabulations of the distributions of these summed number-right scores for the routing and second-stage forms were then evaluated, as described below, to select appropriate routing cut scores for each second-stage form.

The analysis to determine the cut scores included simulations on data from four samples: (1) the fall of kindergarten from the ECLS-K:2011 field test, (2) the fall of kindergarten from the ECLS-K national data collection, (3) the spring of kindergarten from the ECLS-K:2011 field test (estimated), and (4) the spring of kindergarten from the ECLS-K national data collection. For the reading assessment, cut scores were analyzed for the two routing forms (router 1, or the set of EBRS items, and router 2, or the set of additional router items used to determine administration of the middle or high form). Thus, eight simulations were performed for reading, one for each routing form within each sample.

The estimated number of floor and ceiling occurrences also was reviewed using the simulations. To estimate floor effects, the total number of simulated test takers who were predicted to score fewer than three *correct* on the EBRS and low forms was determined. If this number was less than 3 percent of the sample, that would have been taken as evidence of a negligible floor effect. Similarly, if the total number of test takers predicted to score fewer than three *incorrect* on the EBRS, router 2, and high forms was less than 3 percent, that would have been taken as evidence of a negligible ceiling effect. The kindergarten reading simulations showed no evidence of a significant floor or ceiling effect using any of the eight samples. In addition, the counts of simulated test takers who were predicted to have fewer than three *incorrect* on the low form and fewer than three *correct* on the middle form were reviewed to examine whether there was a ceiling effect for the routing/low combination of forms, or a floor effect for the routing/middle combination of forms. Also, the counts of simulated test takers who were predicted to have fewer than three *incorrect* on the middle form and fewer than three *correct* on the high form were reviewed to examine whether there was a floor effect for the routing/high combination of forms, or a ceiling effect for the routing/middle combination of forms. The EBRS, router 2, and low-, middle-, and high-level forms were designed so that each one of them had some items of a similar difficulty level as items included in the other forms, in order to ensure that a child's ability level could still be accurately

measured, if the child was routed to a second-stage form that was not appropriate for that child's ability level.

The approach used to select the optimal cut scores minimized the number of test takers near the cut scores. It also matched the number of students with scores near the lower cut score with the number of students with scores near the upper cut score.

The reading assessment design called for all children to be administered router 1, which consisted of the 20 EBRS items. Analyses were conducted to determine how children proceeding with the remainder of the assessment in English should be routed after completing router 1. This analysis of optimal cut scores indicated that children should be routed directly to the low second-stage form if they had a router 1 score of 9 or lower (including 0) and to the additional routing items in router 2 if they had a score of 10 or higher. Children administered router 2 (20 items) who scored between 0 and 11 items correct on router 2 would proceed with the middle form, while those with scores of 12 or higher would proceed to the high form.

A similar analysis was conducted to determine the mathematics cut scores, the difference being that there was only a single routing form to analyze. This analysis of optimal cut scores indicated that children who responded to 5 or fewer of the 18 items in the router should be routed to the low second-stage form, while children who responded to 6–13 or 14–18 items should be routed to the middle and high forms, respectively.

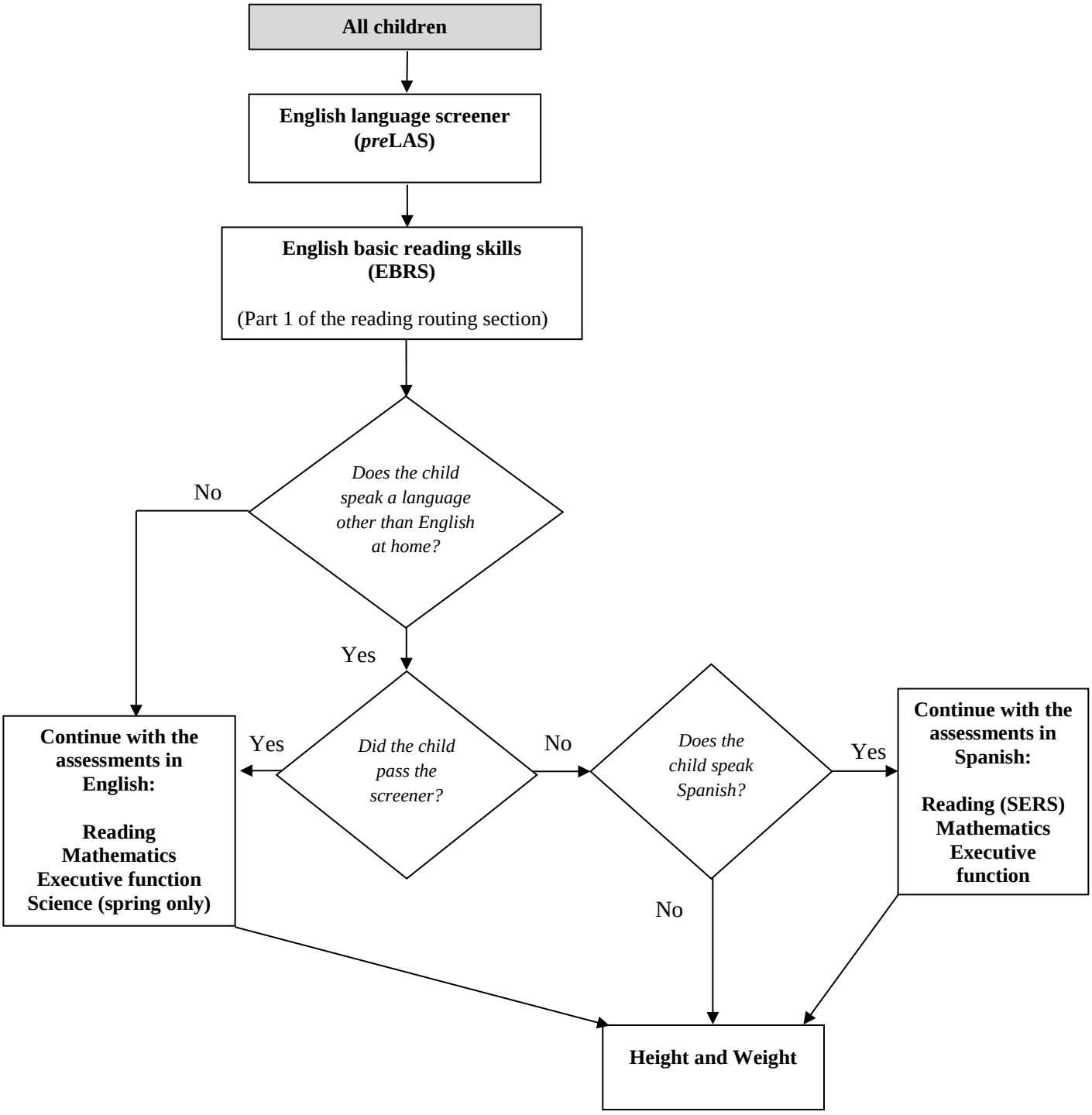
5. PSYCHOMETRIC CHARACTERISTICS OF THE ECLS-K:2011 DIRECT COGNITIVE BATTERY

This chapter documents the results of the direct cognitive assessments for reading, mathematics, and science in the fall 2010 and spring 2011 kindergarten rounds of the ECLS-K:2011. For information about executive function, see chapter 6 and for information on the indirect measures of children's social skills, social relationships, and behavior problems see chapter 7. The emphasis in this chapter is on the psychometric characteristics and scores of the kindergarten academic assessments. Background on the psychometric procedures used to develop and evaluate the scores is provided in chapter 3. The chapter begins with a description in section 5.1 of how children were routed through the direct assessment battery, which is important information for understanding the specific scores that have been developed. Section 5.2 includes the approach to scoring the assessment and types of scores developed. Sections 5.3 through 5.6 focus on the reading, Spanish early reading skills, mathematics, and science assessments, respectively, followed by an evaluation of the longitudinal scale presented in section 5.7. The chapter concludes with a discussion of applications of the scores, including choosing appropriate scores for analysis in section 5.8.

5.1 Routing of Children Through the National Assessment

The full direct assessment battery included assessments in reading, mathematics, science, and executive function, as well as measurements of height and weight, as illustrated in exhibit 5-1. Prior to being administered the reading assessment, all children were administered a language screener, regardless of home language. For children whose primary home language was English, the screener served as a warm-up or practice for the rest of the assessment. While the screener also served as a warm-up for children whose primary home language was other than English, it also determined whether those children understood English well enough to receive the entire direct cognitive assessment battery in English. The screener consisted of two tasks from the *Preschool Language Assessment Scale* (*pre*LAS 2000) (Duncan and De Avila 1998). The "Simon Says" task required children to follow simple, direct instructions given by the assessor in English, such as "point to the floor." The "Art Show" task was a picture vocabulary assessment that tested children's expressive vocabulary. Performance on the ECLS-K:2011 language screener determined which components of the assessment a child received.

Exhibit 5-1. Routing path for the direct child assessment in the ECLS-K:2011 kindergarten year



NOTE: Home language was obtained from school records, the school staff member assigned to coordinate study activities (referred to as the school coordinator), or the child's teacher. Because parents often were not interviewed before children were assessed in school, parent report of home language could not be used to determine assessment routing. SERS = Spanish early reading skills.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

After the language screener, all children moved to the two-stage reading assessment, but the number of items they were administered depended on their performance on the screener. Specifically, the first 20 items of the first-stage routing test, which measured basic reading skills and are therefore referred to collectively as the English basic reading skills (EBRS) items, were administered to all children. The EBRS items target specific early reading skills, predominantly letter recognition and letter sounds, with a few phonemic awareness, vocabulary, and word reading items. Children whose home language was English continued with the rest of the reading assessment after the EBRS, regardless of their performance on the language screener. Children whose home language was not English who achieved at least the minimum score on the language screener also continued with the rest of the reading assessment in English after the EBRS. Children in both of these groups were routed directly to the low-level second-stage test if they did not respond correctly to at least 10 of the 20 EBRS items. If children did respond correctly to at least 10 of the 20 EBRS items, they were administered a second set of 20 routing items (for a total of 40 routing items). Their performance across all 40 items of the routing test determined whether they were administered the middle- or high-level second-stage test (the low-level second-stage test was not considered for these children due to their performance on the EBRS). After the reading assessment, these children were administered the mathematics and executive function assessments, in that order, in both the fall and spring collections. The fall direct assessment battery then ended with measurements of the children's height and weight. In the spring collection, the science assessment was administered between the executive function assessment and the height and weight measurements.

Routing after the EBRS for children whose home language was not English who did not achieve at least the minimum score on the language screener depended on their home language. Spanish-speaking children continued on to the Spanish early reading skills (SERS) assessment, which contained Spanish translations of 31 items from the reading assessment, including 10 items that were also part of the EBRS. Then they were administered the mathematics and executive function assessments that had been translated into Spanish. The science assessment was not translated into Spanish, so after executive function, their height and weight were measured. Children whose primary language was not Spanish were routed out of the rest of the cognitive assessments, and their height and weight were measured.

5.2 Scoring the National Assessment

This section presents information about the assessment scores developed for the kindergarten rounds of data collection, including a discussion of the procedures used to analyze the quality and validity

of the data collected and the scores themselves. Some of the scores are simple counts of correct answers, while others are computed using item response theory (IRT) procedures, which are described in chapter 3. IRT theta and scale scores measure a child's performance on sets of questions with a broad range of difficulty. Raw number-right scores indicate a child's performance with respect to subsets of items.

5.2.1 Confirmation of Unidimensionality

In order to confirm that IRT was an appropriate estimation tool to use for scoring data from each of the assessments, component analyses were run in each domain to determine if indeed the assessment for each domain was measuring a single, dominant component. In both reading and mathematics, the component analyses showed a large single component, but with second and third components that represented higher fractional percentages than expected for unidimensionality. A review of component loadings was performed to identify if the components were representing content components, or simply differences in difficulty levels. For the science and SERS assessments, the component analyses clearly showed a strong single component. Table 5-1 shows the percentage of the variance that each component in each domain explained.

Table 5-1. Component analysis percentages by component by domain, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

	Percentage of Component 1	Percentage of Component 2	Percentage of Component 3
Reading	13.62	8.17	5.14
Mathematics	10.73	6.58	4.43
Science	18.64	6.22	5.52
SERS	30.40	9.18	6.87

NOTE: SERS = Spanish early reading skills.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

In reading, the component structure of the items with loadings was largely consistent with the acquisition of early reading skills in children. The first component primarily included items assessing a student's recognition of letters and the sounds letters make. There are also items at the beginning level of phonemic awareness (i.e., recognizing first sounds in a word), those that probe children's knowledge of and require children to verbally name the letters and produce sounds of letters, and others that assess children's understanding of conventions of print, such as indicating where a word begins. The second

component consisted largely of items assessing student's knowledge of the alphabetic principle, more advanced phonemic awareness (ending sounds of words or manipulating phonemes in words), basic decoding (some blending), and basic, one-syllable word recognition. The skill measured by the items loading on the second component allows the student to read and understand simple sentences. The third component corresponds to items assessing more advanced basic reading ability including recognition of multisyllabic words, sentences containing multisyllabic words, and basic story comprehension. Reading experts reviewed the component structure and, in particular the content and difficulty of the items loading onto each component, and determined that the components seem to correspond relatively well to three phases of reading acquisition sophistication and the associated skills one might expect a child to possess. The items loaded onto components as a function of the item difficulty, rather than the content being measured, which was treated as evidence of the unidimensional construct of reading acquisition.

Similarly, in mathematics, items also loaded onto components as a function of difficulty, not content. Items loading on the first component were those of lower relative difficulty and included basic geometry (that is, knowledge of basic shapes), measurement, number identification, counting, comparing, and ordering numbers. The second component primarily included items from the patterns and algebra content area and those from the data analysis, statistics, and probability content area, while the third component included items from a variety of content categories, all of high difficulty. Number sentences, word problems, and a few items from other categories spanned multiple components, based on item difficulty. As in reading, the component structure indicates multiple components based on increasing item difficulty, not content, and thus validates the assumption of unidimensionality.

5.2.2 Analysis of Differential Item Functioning (DIF)

Before scores were computed for each of the different subject area assessments, an analysis of differential item functioning (DIF) was conducted to determine whether any items should be excluded from scoring because they performed differently for different subgroups of children. (See section 3.4 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.) The Mantel-Haenszel (M-H) and standardized primary item discrepancy index (P-DIF) results agreed for the majority of items, although there were differences in results for some of the items that a high percentage of children answered correctly. Such differences are not unexpected given the nature of the statistical procedures used. Table 5-2 summarizes the results of the M-H DIF and P-DIF analyses for all reading

items for both rounds. Both C-level M-H DIF and P-DIF against one or more race/ethnicity focal groups were observed for 10 items. One item was found to favor the focal group. Upon review, [1] all items were retained for a variety of reasons: bias was not indicated; the item had been previously administered and DIF was not observed; or similar items did not show DIF. DIF procedures were not used for the SERS due to the limited number of children with SERS data in the fall and spring of kindergarten.

Table 5-3 summarizes the results of the DIF analysis of the fall and spring kindergarten rounds combined in mathematics. Five items exhibited both C-level DIF and P-DIF against one or more race/ethnicity focal groups. One item exhibited DIF favoring the focal group. Upon review of the items, no items were removed from scoring since none were determined to exhibit any bias. DIF procedures were also used to analyze the spring kindergarten science assessment. None of the science items exhibited DIF in spring kindergarten.

Table 5-2. Reading assessment differential item functioning, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	4	7	6
Number of DIF items favoring focal group	0	0	0	1

NOTE: The reference group is listed first in each column (i.e., male or White), and the focal group is listed second (i.e., female, Black, Hispanic, or Asian). Reference group cells do not sum to the total number of DIF items for that round 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, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 5-3. Mathematics assessment differential item functioning, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference groups	0	2	3	4
Number of DIF items favoring focal group	0	0	1	0

NOTE: The reference group is listed first in each column (i.e., male or White), and the focal group is listed second (i.e., female, Black, Hispanic, or Asian). Reference group cells do not sum to the total number of DIF items for that round 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, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

[1] Items demonstrating statistical DIF are reviewed by experts from various cultural and ethnic backgrounds to determine if, in fact, the items are exhibiting cultural, ethnic, or sex bias.

5.2.3 Assessment Score Reliability

Estimates of the reliability for each score are computed using the alpha coefficient for the number-right score, and the reliability of the overall IRT ability estimate. Cronbach's alpha, defined as the ratio of the true score to the total score, is used to estimate the internal consistency of the number-right scores. The most appropriate estimate of the reliability of each assessment as a whole is the reliability of the overall IRT ability estimate, theta. This reliability is based on the variance of repeated estimates of theta and applies to theta and 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 cases with scoreable data. The ratio of the within-person variance, averaged over all cases with scoreable data, to the total variance (between-person variance of the posterior mean) is the estimated proportion of total variance that is error variance; 1 minus this proportion is the estimate of true variance, which 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 entire set of scores, rather than estimates evaluated for 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 internal consistency of performance of all items administered and for the full range of variance found in the entire 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. Reliability is a sample-dependent measure of internal consistency of a test and is related to the size of the test. In general, the more items a test has, and the greater the variance in the ability of the test takers, the higher the reliability of the assessment is likely to be.

5.2.4 Item ResponseTheory (IRT)-Based Scores Developed for the ECLS-K:2011

Scores using the full set of assessment items in reading, mathematics, and science were calculated using IRT procedures. As discussed in chapter 3, IRT is a method for modeling assessment data that makes it possible to calculate an overall score for each child that can be compared to scores of other children regardless of which specific items a child is administered. This method was used to calculate scores for the ECLS-K:2011 because, as discussed in chapter 2, the study employed a two-stage assessment (in reading and mathematics) in which children were administered a set of items appropriate for their demonstrated ability level, rather than all the items in the assessment. Although this procedure resulted in children being administered different sets of items, there was a subset of items that all children received (the items in the routing tests, plus a set of items that were administered in more than one of the

different second-stage forms). These common items were used to calculate scores for all children on the same scale. Although in theory all children should have been administered all items in the single-stage assessments (e.g., the science assessment) because there were no discontinue rules or routing into second-stage tests with different items, in practice, not all children have responses for all items in these assessments. Omissions by the child or the discontinuation of the assessment, for example if a child became too tired to continue or refused to answer, resulted in some children who began the single-stage assessments having missing data for some items. In these cases, IRT was used to estimate the child's probability of a correct response when no response information was available. IRT uses the pattern of right, wrong, and omitted responses to the items actually administered in an assessment and the difficulty, discriminating ability,[2] and "guess-ability" of each item to estimate each child's ability on the same continuous scale.

IRT has several advantages over raw number-right scoring. By using the overall pattern of right and wrong responses and the characteristics of each item to estimate ability, IRT can adjust for the possibility of a low-ability child guessing several difficult items correctly. If answers on several easy items are wrong, the probability of a correct answer on a difficult item would be quite low. Omitted items are also less likely to cause distortion of scores, as long as enough items have been answered to establish a consistent pattern of right and wrong answers. Unlike raw number-right scoring, which treats omitted items as if they had been answered incorrectly, IRT procedures use the pattern of responses to estimate the probability of a child providing a correct response for each assessment question. Finally, IRT scoring makes possible longitudinal measurement of gains in achievement, even when the assessments that are administered to a child are not identical at each time point (for example, when a child was administered different levels of the second-stage form in the fall and spring data collections).

5.2.4.1 Theta and the Standard Error of Measurement (*SEM*) of Theta

The theta score is an estimate of a child's ability in a particular domain (e.g., reading, mathematics, or science) based on that child's performance on the items administered. This score represents a child's latent ability and is not dependent on the difficulty of the items a child was administered. Theta scores for reading and mathematics are developed for both the fall kindergarten and spring kindergarten data collection rounds. Theta scores for science are available only for the spring

[2] The discriminating ability describes how well changes in ability level predict changes in the probability of answering the item correctly at a particular ability level.

kindergarten round since the science assessment was not administered in the fall of kindergarten. The theta scores are reported on a metric ranging from -6 to 6, with lower scores indicating lower ability and higher scores indicating higher ability.

Gain scores in each domain may be obtained by subtracting the IRT thetas at an earlier administration from the IRT thetas at a later administration. Thetas for different subject areas are not comparable to each other because scores are calibrated separately within each domain (for example, if a child's IRT theta in reading is higher than in mathematics, it would not be appropriate to interpret that to mean the child is doing better in reading than in mathematics).

The estimated standard error of theta provides a measure of uncertainty of the theta score estimate for each child. Adding and subtracting twice the standard error estimate from the theta score estimates provides an approximate 95 percent confidence interval or range of values that is likely to include the child's true theta score. Unlike classical item theory, where the precision of the scores is usually assumed to be consistent across all examinees, IRT procedures usually provide an estimate of the accuracy of the theta estimate for each test taker. Measurements are most accurate for test takers who answer relatively more questions with a difficulty that is close to their ability level. As discussed in chapter 4, each subject area assessment was designed with the difficulty of most of the test items spaced across a range defined by plus or minus two standard deviations of the expected average theta. There were relatively fewer items administered in the tails beyond two standard deviations; therefore, children at the extremes of the ability range received relatively fewer items matched to their ability level and, therefore, their estimated standard errors of measurement can be expected to be greater.

5.2.4.2 IRT Scale Scores

The IRT-based overall scale score for each content domain is an estimate of the number of items a child would have answered correctly in each data collection round if that child had been administered all of the questions for that domain. To calculate the IRT-based overall scale score for each domain, a child's theta is used to predict a probability for each assessment item that the child would have gotten the item correct. Then, the probabilities for all the items administered as part of the domain (i.e., reading, math, or science) are summed to create the overall scale score. Because the computed scale scores are sums of probabilities, the scores are not integers.

The probability that a child would have gotten an item correct is dependent on the difficulty, discrimination, and guessing parameter estimates of the item, as well as the ability estimate (theta) of the child. For example, in an item set designed for both fall and spring administrations, where some items have high difficulty parameter estimates to target the expected ability levels in spring, the predicted probability that an average child would answer each of those high difficulty items correctly in the fall would be low, resulting in average scale scores that are lower in the fall than in the spring. As a result, the distribution of scale scores can be skewed.

As with the IRT thetas, gain scores in each domain may be obtained by subtracting the IRT scale scores at an earlier administration from the IRT scale scores at a later administration. It is important to note again that scores for different subject areas are not comparable to each other and that it would not be appropriate to interpret scores that are higher in one domain to mean the child is doing better in that domain than in another.

5.2.5 Raw Number-Right Scores for the ECLS-K:2011

Several raw number-right scores, which are counts of the number of items a child answered correctly, are provided on the data file for the kindergarten rounds of data collection. Raw number-right scores for the Simon Says and Art Show subtests of the *pre*LAS provide information on receptive and expressive vocabulary components of children's basic English proficiency. They are derived from the 10 Simon Says items and the 10 Art Show items. The Simon Says and Art Show subtests of the *pre*LAS were administered to all children, so all children have raw number-right scores for these two subtests.

A raw number-right score also is provided for children's performance on the set of 20 EBRS items, which were administered to all children as part of the reading assessment routing test. Additionally, for those children who were routed to the SERS, number-right scores are provided for the 10 items common to the EBRS and SERS.

5.3 Reading Assessment

5.3.1 Samples and Associated Statistics

The kindergarten reading assessment consisted of 40 routing items (20 items in part 1 of the router, referred to as the EBRS, and 20 items in part 2 of the router), followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 17, 14, and 21 items, respectively. Discontinue rules were employed within the second-stage tests of the 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 reading assessment and the assessment's associated statistics are shown in table 5-4. There was no evidence of floor or ceiling effects (based on the low numbers of children with chance or perfect scores, respectively) on the reading assessment during the fall or spring kindergarten rounds. The number of children with scoreable data is defined as the number of children who responded (correctly or incorrectly) to at least 10 items in the domain. In addition, review of the classical item analysis *r*-biserials (see section 3.1 for details) in both fall and spring showed two items with slightly lower *r*-biserials than the ideal minimum of 0.3: one item that was very difficult for the majority of the sample, and one that was very easy for the majority of the sample.

Table 5-4. Kindergarten reading assessment samples, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Characteristic	Fall kindergarten		Spring kindergarten	
	Number	Percent	Number	Percent
Total sample size	16,410	100	17,810	100
Number with no reading items, or fewer than 10	120	1	30	#
Number of children with scoreable data	16,290	99	17,770	100
Number and percent of children with responses only for the router, first 20 items (EBRS)	350	2	180	1
Number and percent of children with responses only for the router, all 40 items	10	#	#	#
Number and percent of children routed to low form	3,110	19	430	2
Number and percent of children routed to middle form	11,460	70	10,290	58
Number and percent of children routed to high form	1,360	8	6,880	39
Number and percent of children with a perfect score: router + high form	#	#	#	#
Number and percent of children with a chance score or below: router, first 20 items (EBRS) + low form	10	#	#	#

Rounds to zero.
NOTE: Estimates are based on the children assessed in English. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Unweighted sample sizes are rounded to the nearest 10. Percentages are unweighted. Percentages and sample sizes may not sum to total due to rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2010 and spring 2011.

5.3.2 Score Statistics

Table 5-5 presents summary statistics for the IRT-based reading scores including the reading theta, the standard error of measurement (*SEM*) of theta, and the IRT scale scores, which indicate predicted performance on the 83 unique items administered in the kindergarten rounds. Table 5-6 presents summary statistics for the raw number-right (i.e., non-IRT-based) reading scores, which indicate performance on the two *pre*LAS tasks administered in the language screener and on the EBRS items (the first 20 items of the router). These raw number-right scores are integers based on the total number of

items administered in each subset. Both the IRT-based scores and the raw number-right scores are calculated for all children with scoreable reading assessment data.

Table 5-5. Reading assessment statistics, by IRT-based score, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Variable	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1RTHET	X1 Reading IRT theta score	15,670	-6.0–+6.0	-0.57	0.871
X2RTHET	X2 Reading IRT theta score	17,190	-6.0–+6.0	0.48	0.772
X1RSETH	X1 Reading IRT *SEM* of theta	15,670	0.0–+6.0	0.29	0.044
X2RSETH	X2 Reading IRT *SEM* of theta	17,190	0.0–+6.0	0.24	0.067
X1RSCAL	X1 Reading IRT scale score	15,670	0.0–+83.0	34.42	11.663
X2RSCAL	X2 Reading IRT scale score	17,190	0.0–+83.0	49.08	11.724

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10.
IRT = item response theory. *SEM* = standard error of measurement.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 5-6. Reading assessment statistics, by raw number-right score, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Variable	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1PLSS	X1 *pre*LAS Simon Says raw number-right score	15,780	0–10	9.18	1.754
X2PLSS	X2 *pre*LAS Simon Says raw number-right score	17,220	0–10	9.60	1.120
X1PLART	X1 *pre*LAS Art Show raw number-right score	15,780	0–10	9.26	1.705
X2PLART	X2 *pre*LAS Art Show raw number-right score	17,220	0–10	9.54	1.274
X1PLTOT	X1 *pre*LAS total raw number-right score	15,780	0–20	18.43	3.184
X2PLTOT	X2 *pre*LAS total raw number-right score	17,220	0–20	19.14	2.178
X1EBRSTOT	X1 EBRS raw number-right score	15,740	0–20	13.18	4.424
X2EBRSTOT	X2 EBRS raw number-right score	17,200	0–20	17.06	2.976

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10.
IRT = item response theory. *SEM* = standard error of measurement.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

5.3.3 Reliabilities

Table 5-7 presents the reliability statistics for the reading assessment scores. The reliabilities shown in table 5-7 are typical and adequate for tests with these numbers of items.

Table 5-7. Reading assessment reliabilities, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Score	Number of items	*n*	Reliability
Fall kindergarten			
IRT-based scores	83	15,670	.95
*pre*LAS Simon Says raw number-right score	10	15,780	.85
*pre*LAS Art Show raw number-right score	10	15,780	.86
*pre*LAS total raw number-right score	20	15,780	.91
EBRS raw number-right score	20	15,740	.87
Spring kindergarten			
IRT-based scores	83	17,190	.95
*pre*LAS Simon Says raw number-right score	10	17,220	.79
*pre*LAS Art Show raw number-right score	10	17,220	.82
*pre*LAS total raw number-right score	20	17,220	.89
EBRS raw number-right score	20	17,190	.97

NOTE: 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 unweighted sample size is the number of cases with a valid score rounded to the nearest 10. IRT = item response theory. EBRS = early basic reading skills.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

5.4 Spanish Early Reading Skills (SERS)

5.4.1 Samples and Associated Statistics

As mentioned in section 5.1, Spanish-speaking children who did not achieve at least the minimum score on the *pre*LAS subtests that made up the language screener were administered the SERS items. Therefore, scores for the SERS are only available for these children. IRT-based statistics indicating performance on the 31 SERS items include theta, the standard error of measurement (*SEM*) of theta, and the IRT-based scale score. Additionally, raw number-right scores are available for the 10 items that were administered in both English (as part of the EBRS) and in Spanish (as part of the SERS). The samples and associated statistics for the SERS are shown in table 5-8. There was no evidence of floor or ceiling effects (based on the low number of children with chance or perfect scores, respectively) on the SERS

assessment during the fall or spring kindergarten rounds. Review of the *r*-biserials in both fall and spring showed two items with slightly lower *r*-biserials than ideal; one item that was very difficult and one that was very easy for the majority of the sample.

Table 5-8. Kindergarten SERS assessment samples, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

	Fall kindergarten		Spring kindergarten	
Characteristics	Number	Percent	Number	Percent
Total sample size	320	100	150	100
Number and percent of children with a perfect score	0	0	#	1
Number and percent of children with a chance score	10	2	#	1

Rounds to zero.
NOTE: Estimates are based on the children assessed in Spanish. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Percentages are unweighted. SERS = Spanish early reading skills.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

5.4.2 Score Statistics

Table 5-9 presents summary statistics for the IRT-based SERS scores including theta, the standard error of measurement of theta *(SEM)*, and the IRT scale scores, which indicate predicted performance on the 31 unique items administered in the kindergarten rounds. Table 5-10 presents summary statistics for the raw number-right (i.e., non-IRT-based) SERS scores, which indicate performance on the set of 10 common items administered in English in the EBRS, and administered in Spanish in the SERS. These raw number-right scores are integers based on the 10 items administered in each subset. Both the IRT-based scores and the raw number-right scores are calculated for all children with scoreable assessment data.

Table 5-9. Kindergarten SERS assessment statistics, by IRT-based score, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Variable	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1SERSTH	X1 SERS IRT theta score	310	-6.0–+6.0	-0.41	0.852
X2SERSTH	X2 SERS IRT theta score	150	-6.0–+6.0	0.68	0.667
X1SERSSE	X1 SERS IRT *SEM* of theta	310	0.0–6.0	0.39	0.142
X2SERSSE	X2 SERS IRT *SEM* of theta	150	0.0–6.0	0.27	0.077
X1SERSSC	X1 SERS IRT scale score	310	0.0–31.0	12.75	5.343
X2SERSSC	X2 SERS IRT scale score	150	0.0–31.0	20.83	5.592

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10.
IRT = item response theory. *SEM* = standard error of measurement. SERS = Spanish early reading skills.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 5-10. Kindergarten SERS assessment statistics, by raw number-right score, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Variable	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1EBRSCM	X1 EBRS common raw number-right score, EBRS	340	0–10	3.21	2.583
X2EBRSCM	X2 EBRS common raw number-right score, EBRS	150	0–10	4.13	3.129
X1SERSCM	X1 SERS common raw number-right score, SERS	320	0–10	4.72	2.992
X2SERSCM	X2 SERS common raw number-right score, SERS	150	0–10	8.05	2.037

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. EBRS = early basic reading skills. SERS = Spanish early reading skills.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

5.4.3 Reliabilities

Table 5-11 presents the reliability statistics for the scores of the SERS assessment for the kindergarten rounds. All scores, except the EBRS common raw number-right score, have reliabilities, which are typical and adequate for assessments of this many items. The EBRS common raw number-right score has a lower reliability than generally acceptable due to the confluence of the few numbers of items in the score with the lack of variability in the Spanish-speaking children's abilities in English.

Table 5-11. Kindergarten SERS assessment reliabilities, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Score	Number of items	*n*	Reliability
Fall kindergarten			
IRT-based scores	31	310	.99
EBRS common raw number-right score	10	340	.80
SERS common raw number-right score	10	320	.87
Spring kindergarten			
IRT-based scores	31	150	.99
EBRS common raw number-right score	10	150	.69
SERS common raw number-right score	10	150	.84

NOTE: 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 unweighted *n* is the rounded number of cases with a valid score.
IRT = item response theory. EBRS = early basic reading skills. SERS = Spanish early reading skills.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

5.5 Mathematics Assessment

5.5.1 Samples and Associated Statistics

The kindergarten mathematics assessment consisted of 18 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 17, 23, and 28 items, respectively. Table 5-12 displays the total number of children administered the mathematics assessment and the assessment's associated statistics for the fall and spring kindergarten rounds. No significant ceiling or floor effects (i.e., based on the low numbers of children with perfect or chance scores, respectively) were observed in the fall or spring kindergarten mathematics tests. Classical item analysis results for the English mathematics administration showed two items with *r*-biserials lower than ideal: one very easy and one very difficult item for the sample. The Spanish-translated mathematics assessment showed more items (five) with less than ideal *r*-biserials, a result of the low number of observations, suggesting the *r*-biserial calculations for this small subset may be unreliable.

Table 5-12. Kindergarten mathematics assessment samples, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Characteristics	Fall kindergarten		Spring kindergarten	
	Number	Percent	Number	Percent
Total sample size	16,730	100	17,960	100
Number with no mathematics items, or fewer than 10	510	3	220	1
Number of children with scoreable data	16,210	97	17,730	99
Number and percent of children with responses only for the router	0	0	#	#
Number and percent of children routed to low form	4,220	26	960	5
Number and percent of children routed to middle form	10,310	64	9,690	55
Number and percent of children routed to high form	1,690	10	7,090	40
Number and percent of children with a perfect score: router + high form	#	#	#	#
Number and percent of children with chance score or below: router + low form	40	#	10	#

Rounds to zero.
NOTE: Estimates are based on the children assessed in English or Spanish. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Unweighted sample sizes are rounded to the nearest 10. Percentages are unweighted. Percentages and sample sizes may not sum to total due to rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2010 and spring 2011.

5.5.2 Score Statistics

Only IRT-based scores were produced for the mathematics assessment. Table 5-13 presents summary statistics for the mathematics thetas, the standard errors of measurement (*SEM*) of thetas, and the IRT scale scores, using the 75 unique items administered in the kindergarten rounds.

Table 5-13. Mathematics assessment statistics, by IRT-based score, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Variable	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1MTHET	X1 Math IRT theta score	15,600	-6.0–+6.0	-0.52	0.929
X2MTHET	X2 Math IRT theta score	17,140	-6.0–+6.0	0.42	0.773
X1MSETH	X1 Math IRT *SEM* of theta	15,600	0.0–6.0	0.35	0.101
X2MSETH	X2 Math IRT *SEM* of theta	17,140	0.0–6.0	0.28	0.064
X1MSCAL	X1 Math IRT scale score	15,600	0.0–75.0	28.95	10.694
X2MSCAL	X2 Math IRT scale score	17,140	0.0–75.0	41.64	11.166

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10.
IRT = item response theory. *SEM* = standard error of measurement.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

5.5.3 Reliabilities

Table 5-14 presents reliability statistics for the scores of the fall and spring kindergarten mathematics assessments (calculated in the same way as the reading reliability statistics, described in section 5.3.3). The reliabilities shown in table 5-14 are typical and adequate for this test with this number of items.

Table 5-14. Mathematics assessment reliabilities, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Score	Number of items	*n*	Reliability
Fall kindergarten			
IRT-based scores	75	15,600	0.92
Spring kindergarten			
IRT-based scores	75	17,140	0.94

NOTE: 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 unweighted sample size is the number of cases with a valid score rounded to the nearest 10. IRT = item response theory.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

5.5.4 Comparability of Spanish Mathematics Test

The mathematics assessment was translated from English into Spanish. It was administered in English to all children who achieved at least a minimum score on the language screener, regardless of home language, and administered in Spanish to Spanish-speaking children who did not achieve at least a minimum score on the language screener. DIF procedures were used to determine whether the mathematics items performed similarly regardless of the language of administration. Assessment items showing DIF by the language of administration would be an indication that children with equal mathematics ability demonstrated differential performance attributable to the language in which the item was administered. The lack of DIF would indicate that children with equal mathematics ability demonstrated similar performance regardless of the language in which the item was administered, meaning the assessments were indeed comparable and pooling of the samples for IRT calibration would be appropriate.

For this DIF analysis, the mathematics assessment item data for children who were administered the English-language version (reference group) were compared with assessment item data for children who were administered the Spanish-language version. The results showed that none of the items exhibited significant DIF, as defined in section 3.4.[3] Given the lack of DIF observed between the items when administered in English or Spanish, data from both administrations were calibrated and scored as the same assessment, regardless of the language of administration, rather than scoring them as separate assessments.

Table 5-15 shows the samples and associated statistics of the mathematics assessment by language of administration.

[3] Although a minimum of 500 cases is generally required for DIF analyses, the English-language and Spanish-language comparison sample sizes did not sum to 500, even in the fall and spring combined. With smaller sample sizes, artificial statistical DIF may result, which is why the minimum sample size is required. Thus, with the lower number of observations, and still the non-existence of DIF, the analyses are considered valid.

Table 5-15. Kindergarten mathematics assessment samples, by language of assessment, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

	Fall kindergarten				Spring kindergarten			
	English		Spanish		English		Spanish	
Characteristic	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Total sample size	16,100	100	320	100	17,810	100	150	100
Number of children with no mathematical items or fewer than 10	200	3	#	1	220	1	#	1
Number of children with scoreable data	15,900	97	312	99	17,580	99	150	99

Rounds to zero.
NOTE: Unweighted sample sizes are rounded to the nearest 10. Percentages are unweighted. Percentages and sample sizes may not sum to total due to rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2010 and spring 2011.

5.6 Science Assessment

5.6.1 Samples and Associated Statistics

The spring kindergarten science assessment included a total of 20 items administered to all children who were routed to the assessment. Table 5-16 shows the total number of children who were administered the science assessment and the assessment's associated statistics for the spring kindergarten round. No significant ceiling or floor effects (i.e., based on the low numbers of children with perfect or chance scores, respectively) were observed in the spring kindergarten science test. The classical item analysis results showed one difficult item with an *r*-biserial slightly lower than ideal.

Table 5-16. Kindergarten science assessment sample: Spring 2011

Characteristics	Number	Percent
Total sample size	17,810	100
Total with no science items, or fewer than 10	280	2
Number of children with scoreable data	17,520	98
Number and percent of children with a perfect score	20	#
Number and percent of children with chance score or below	270	2

Rounds to zero.
NOTE: Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Unweighted sample sizes are rounded to the nearest 10. Percentages are unweighted. Percentages and sample sizes may not sum to total due to rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), spring 2011.

5.6.2 Score Statistics

Only IRT-based scores were produced for the science assessment. Table 5-17 presents summary statistics for the science theta, the *SEM* of theta, and the scale score, which indicate performance on the 20 items administered in the spring kindergarten round.

Table 5-17. Science assessment statistics, by IRT-based score, ECLS-K:2011 spring kindergarten data collection: Spring 2011

Variable	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X2STHET	X2 Science IRT theta score	16,940	-6.0–+6.0	0.00	0.887
X2SSETH	X2 Science IRT *SEM* of theta	16,940	0.0–6.0	0.70	0.081
X2SSCAL	X2 Science IRT scale score	16,940	0.0–20.0	11.22	2.861

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10.
IRT = item response theory. *SEM* = standard error of measurement.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2011.

5.6.3 Reliabilities

Table 5-18 presents reliability for the spring kindergarten science assessment scores (calculated in the same way as the reading reliability statistics, described in section 5.3.3). As noted

above, 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. Therefore, relative to the reading and mathematics assessments, which had more items, the lower reliability of the IRT-based scores from the spring science assessment is expected.

Table 5-18. Science assessment reliability, ECLS-K:2011 spring kindergarten data collection: Spring 2011

Score	Number of items	*n*	Reliability
IRT-based scores	20	16,940	0.75

NOTE: 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 unweighted sample size is the number of cases with a valid score rounded to the nearest 10. IRT = item response theory.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2011.

5.7 Applications

This section provides guidance in the selection and use of scores for analyzing status and gain in cognitive knowledge and skills.

5.7.1 Choosing the Appropriate Score for Analysis

When choosing scores to use in analysis, researchers should consider the nature of their research questions, the type of statistical analysis to be conducted, the population of interest, and the audience for their research findings. The sections below discuss the general suitability of the different types of scores for different analyses.

- The IRT-based theta scores are overall measures of ability. They are appropriate for both cross-sectional and longitudinal analyses. They are useful in examining differences in overall achievement among subgroups of children in a given data collection round or across rounds, as well as in analysis looking at correlations between achievement and child, family, and school characteristics. The fall kindergarten and spring kindergarten theta scores are on the same metric. Therefore, an analyst looking at growth across the kindergarten year could subtract the fall kindergarten score from the spring kindergarten score to compute a gain score. The theta scores may be more desirable than the scale scores for use in a multivariate analysis because their distribution generally tends to be more normal than the

distribution of the scale scores.[4] However, for a broader audience of readers unfamiliar with IRT modeling techniques, the metric of the theta scores (from -6 to 6) may be less readily interpretable than the metric of the scale scores. Researchers should consider their analysis and the audience for their research when selecting between the theta and the scale score.

- The IRT-based scale scores also are overall measures of achievement. They are appropriate for both cross-sectional and longitudinal analyses. They are useful in examining differences in overall achievement among subgroups of children in a given data collection round or across rounds, as well as in analysis looking at correlations between achievement and child, family, and school characteristics. The fall kindergarten and spring kindergarten scale scores are on the same metric. Therefore, an analyst looking at growth across the kindergarten year could subtract the fall kindergarten score from the spring kindergarten score to compute a gain score. Results expressed in terms of scale score points, scale score gains, or an average scale score may be more easily interpretable by a wider audience than results based on the theta scores.

- *pre*LAS subtest raw number-right scores provide information on children's basic English proficiency. These scores may be of interest to users conducting research on children with limited English proficiency. However, because of the limited number of items included in these subtests, these scores do not represent a comprehensive measure of proficiency or of reading skills and knowledge. The primary purpose of fielding these subtests in the ECLS-K:2011 was so they could be used as an English language proficiency screener. The majority of children in the ECLS-K:2011 scored highly or near perfect on these subtests, which was expected given that the subtests came from a standardized assessment for preschoolers and the majority of ECLS-K:2011 children spoke English, even if it was not their primary home language. The *pre*LAS scores are of limited value for children who were not English language learners. The IRT-based reading theta or scale scores, which are available for all children, should be used by analysts interested in performance on the reading assessment, regardless of a child's home language.

- EBRS raw number-right scores provide information on children's performance on the first 20 items administered to all children as part of the reading assessment routing test. These EBRS scores would be useful for someone with a specific analytic interest in the knowledge and skills covered in this particular item set, which are among the most basic knowledge and skills measured in the reading assessment. As with the *pre*LAS subtest items, children who were not English language learners tended to do well on this section of the assessment, and so these scores may be of limited value for analysis of their knowledge and skills. Also, since these are raw scores, the difficulty of the items children answered correctly is not reflected in the score. A child who answered only the first 10 items correctly would have the same score as a child who answered 5 easier and 5 more difficult items correctly. The IRT-based reading theta or scale scores, which are available for all children, should be used by analysts interested

[4] It is recommended that analysts review the distributions for normality. In assessments where the number of items or number of observations is low, normality of distributions may be affected. In the ECLS-K:2011, both the science and SERS distributions departed from normal, due to the limited number of items and observations, respectively.

in overall performance on the reading assessment, regardless of a child's home language.

- EBRS/SERS common item raw number-right scores provide information on Spanish-speaking children's performance on 10 items that were administered in both English and Spanish. Researchers may find these scores useful in an analysis focusing on Spanish-speaking children with limited English proficiency because the scores allow for a comparison of the number of correct responses in English with the number of correct responses in the child's primary home language. It is important to note that these items are direct translations from the existing English items to Spanish. They have not been scaled together, and the item difficulties may not be exactly comparable from one language to the other. Although this is the case, the items have very limited language load, and expert reviewers selected items that translated easily and could be expected to be roughly equivalent in difficulty in either language.

5.7.2 Analytic Considerations for Measuring Gains in the ECLS-K:2011

An important issue to be considered when analyzing achievement scores and gains is assessment timing: children's age at assessment, the date of assessment, and the time interval between assessments. Most sampled children were born throughout the second half of 2004 and first half of 2005, but their birth dates were not related to testing dates. As a result, children were tested at different developmental and chronological ages. Assessment dates ranged from August to December for the fall data collection, and from January to July for the spring round. Children assessed later in a data collection period, for example in December of the fall collection, may be expected to have an advantage over children assessed earlier in the data collection period, for example in the first days or weeks of school, because they had more exposure to educational content before being assessed. Substantial differences in the intervals between assessments may also affect analysis of gain scores. Children assessed in September for the fall data collection and June for the spring data collection have more time to learn skills than children assessed in November and March. These differences in intervals may or may not have a significant impact on analysis results. In designing an analysis plan, it is important to consider whether and how differences in age, assessment date, and interval may affect the results; to look at relationships between these factors and other variables of interest; and to adjust for differences, if necessary.

When using the IRT scale scores as longitudinal measures of overall growth, analysts should keep in mind that gains made at different points on the scale have qualitatively different interpretations. Children who made gains toward the lower end of the scale, for example, in skills such as identifying letters and associating letters with sounds, are learning different skills than children who made gains at

the higher end of the scale, for example, those who have gone from reading single words to reading sentences, although their gains in number of scale score points may be the same. Comparison of gains in scale score points is most meaningful for groups that started with similar initial status. One way to account for children's initial status is to include a prior round assessment score as a control variable in an analytic model. For example, the fall scale score could be included in a model using the spring scale score as the outcome.

6. PSYCHOMETRIC CHARACTERISTICS OF THE EXECUTIVE FUNCTION MEASURES

Executive functions are interdependent processes that work together to regulate and orchestrate cognition, emotion, and behavior and that help a student to learn in the classroom (e.g., Diamond 2013). Measures of executive function were included in the kindergarten direct child assessment battery to assess children's cognitive flexibility and working memory: the Dimensional Change Card Sort (DCCS) and the Numbers Reversed subtest of the *Woodcock-Johnson III Tests of Cognitive Abilities,* respectively.

6.1 Dimensional Change Card Sort (DCCS)

The Dimensional Change Card Sort (Zelazo 2006) is used to collect information on children's cognitive flexibility. In this task, children are asked to sort a series of 22 picture cards according to different rules. Each card has a picture of either a red rabbit or a blue boat. The children are asked to sort each card into one of two trays depending on the sorting rule they have been told to use. One tray has a picture of a red boat, and the other has a picture of a blue rabbit. For the first set of items, the Color Game (each set is referred to as a game), the rule is to sort the cards by color (i.e., red or blue). For example, a blue boat card would be sorted into the blue rabbit tray. In the second game, the Shape Game, the rule is to sort the cards by shape (i.e., rabbit or boat). For example, a red rabbit card would be sorted into the blue rabbit tray. If the child correctly sorts four of the six cards in the Shape Game, then he or she moves on to the third game: the Border Game. In the Border Game, the sorting rule (by color or by shape) depends on whether or not the card has a black border around the edges. If the card has a border, the child is to sort by color; if there is no border on the card, the child is to sort by shape.

Item-level data for the Dimensional Change Card Sort are provided on the base-year data file. There are six variables with results for the Color Game, six variables with results for the Shape Game, and six variables with results for the Border Game. There were four practice items administered to the children, but the results from these practice items are not included in the data file. The item-level data for the Color and Shape games are scored "Correct" (i.e., card sorted into the correct tray according to the sorting rule) or "Incorrect" (i.e., card sorted into the incorrect tray). There is a third score provided for the Border Game, "Not administered"; this code indicates that the child was not administered the item because he or she did not answer enough items correctly to advance to this item in the assessment. The

"Not administered" code is different than a system missing code in that only those children who were administered the Dimensional Change Card Sort could have a "Not administered" code. If a child was not administered the Dimensional Change Card Sort at all, that case would have a missing code for the Dimensional Change Card Sort scores. Variable names for the item-level data from the fall kindergarten assessments begin with "C1," and the variable names for the item-level data from the spring kindergarten assessments begin with "C2." The Dimensional Change Card Sort was administered in Spanish for children routed through the assessment in Spanish. Data from the English and Spanish administrations are combined into the same item-level variables. Researchers who want to account for language of administration in their analyses can use the variables X1FLSCRN and X2FLSCRN, which are also in the data file, to identify which cases were administered the Dimensional Change Card Sort in English and which cases were administered it in Spanish.

6.1.1 Mean Scores

Using scoring rules provided by the developers, two scores were developed from the Dimensional Change Card Sort data collected in the kindergarten rounds of data collection: the post-switch score and the Border Game score. The post-switch score is the number of cards the child correctly sorted by shape (i.e., after switching from sorting by color to sorting by shape), not including the practice trials. The Border Game score is the number of cards the child correctly sorted when the sorting rule was determined by the presence (or absence) of a border around the card.

Table 6-1 presents the variable names, descriptions, value ranges, weighted means, and standard deviations for the post-switch and Border Game scores for the fall kindergarten and spring kindergarten Dimensional Change Card Sort. Children who did not correctly sort at least four of the six cards in the Shape Game were not administered the Border Game and do not have a Border Game score. As a result, the *n* with valid (i.e., nonmissing) data for the post-switch score is different than the *n* with valid (i.e., nonmissing) data for the Border Game score. For more information on the administration procedures and the scores for the Dimensional Change Card Sort, see *The Dimensional Change Card Sort (DCCS): A Method of Assessing Executive Function in Children* (Zelazo 2006).

Table 6-1. Dimensional Change Card Sort variable names, descriptions, value ranges, weighted means, and standard deviations: School year 2010–11

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1CSPSSC	X1 Card Sort post-switch score	15,600	0–6	5.23	1.679
X1CSBGSC	X1 Card Sort Border Game score	13,280	0–6	3.70	1.185
X2CSPSSC	X2 Card Sort post-switch score	17,150	0–6	5.55	1.210
X2CSBGSC	X2 Card Sort Border Game score	15,690	0–6	4.10	1.314

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

The post-switch score has a relatively high mean (table 6-1), indicating that the majority of children did well on the portion that asked them to sort by shape. According to the Dimensional Change Card Sort developer, given this pattern in the data, researchers should create a single Dimensional Change Card Sort composite score by summing the post-switch score and the Border Game score and use that combined score in analyses. Before creating this combined score, researchers should be sure to recode the reserve codes appropriately (for more information on the reserve, or missing value, codes used in the data file, please see section 7.3 of *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11, User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic Codebook, Public Version* (NCES 2015-074) [Tourangeau et al. 2015]). When recoding reserve codes, inapplicable (-1) codes on the Border Game should be recoded as 0, because an inapplicable code on the Border Game score indicates that a child did not answer enough Shape Game (i.e., post-switch) items correctly to advance to the Border Game, and therefore "answered" 0 of the Border Game items correctly.

Mean Dimensional Change Card Sort scores by data collection round and child characteristics are presented in tables 6-2 and 6-3. As can be seen in table 6-2, there is little variation in scores for the Dimensional Change Card Sort post-switch score. Approximately 74 percent of the children administered the Card Sort achieved 6 out of 6 correct in fall kindergarten, and approximately 80 percent of the children achieved 6 out of 6 correct in spring kindergarten. This finding is consistent with expectations. Zelazo (2006) found that by age 5 most children are able to switch sorting rules when asked to do so. In the ECLS-K:2011, the mean age at assessment was 5 years, 8 months for the fall kindergarten round and 6 years, 2 months for the spring kindergarten round. The Border Game is administered to add difficulty for children who can successfully complete the post-switch items (i.e., who achieve a score 5 of 6 on the Shape Game).

Table 6-2. Mean Dimensional Change Card Sort post-switch score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	15,600	5.23	1.679	17,150	5.55	1.210
Sex						
Male	7,930	5.16	1.743	8,710	5.50	1.277
Female	7,610	5.30	1.604	8,370	5.61	1.128
Race/ethnicity						
White, non-Hispanic	7,590	5.49	1.275	8,090	5.68	0.928
Black, non-Hispanic	2,110	4.86	2.064	2,240	5.30	1.566
Hispanic	3,760	4.86	2.063	4,300	5.42	1.450
Asian, non-Hispanic	1,140	5.23	1.670	1,430	5.55	1.223
Hawaiian, Other Pacific Islander, non-Hispanic	80	4.77	2.118	110	5.05	1.944
American Indian/Alaska Native, non-Hispanic	150	5.03	1.756	160	5.65	1.234
More than one race, non-Hispanic	700	5.44	1.365	750	5.62	1.053
School type						
Public school	13,530	5.21	1.709	15,000	5.54	1.237
Private school	2,080	5.36	1.414	2,130	5.67	0.958

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. The range of possible values is 0 to 6.Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 6-3. Mean Dimensional Change Card Sort Border Game score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	13,280	3.70	1.185	15,690	4.10	1.314
Sex						
Male	6,640	3.68	1.188	7,860	4.07	1.316
Female	6,590	3.72	1.182	7,770	4.13	1.312
Race/ethnicity						
White, non-Hispanic	6,830	3.82	1.243	7,600	4.28	1.330
Black, non-Hispanic	1,660	3.50	1.050	1,950	3.73	1.185
Hispanic	2,940	3.48	1.065	3,830	3.84	1.260
Asian, non-Hispanic	980	3.77	1.224	1,300	4.16	1.329
Hawaiian, Other Pacific Islander, non-Hispanic	60	3.65	0.976	100	3.77	1.216
American Indian/Alaska Native, non-Hispanic	130	3.83	0.945	150	4.19	1.208
More than one race, non-Hispanic	620	3.77	1.182	690	4.26	1.338
School type						
Public school	11,480	3.68	1.175	13,680	4.08	1.312
Private school	1,800	3.87	1.244	1,990	4.26	1.317

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. The range of possible values is 0 to 6. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

6.2 Numbers Reversed

This measure assesses the child's working memory. It is a backward digit span task that requires the child to repeat an orally presented sequence of numbers in the reverse order in which the numbers are presented. For example, if presented with the sequence "3…5," the child would be expected to say "5…3." Children are given 5 two-number sequences. If the child gets 3 consecutive two-number sequences incorrect, then the Numbers Reversed task ends. If the child does not get 3 consecutive two-number sequences incorrect, the child is then given 5 three-number sequences. The sequence becomes increasingly longer, up to a maximum of eight numbers, until the child gets three consecutive number sequences of the same length incorrect (or completes all number sequences).

Item-level data for the Numbers Reversed subtask are provided in the base-year data file. The maximum number of items any child was administered in either the fall or spring kindergarten collections was 30 items (5 two-digit number items; 5 three-digit number items; 4 four-digit number items; 4 five-digit number items; 4 six-digit number items; 4 seven-digit number items; and 4 eight-digit number items). Each item is scored "Correct" (i.e., the child correctly repeated the number sequence in reversed order), "Incorrect" (i.e., the child did not correctly repeat the number sequence in reversed order), or "Not administered" (i.e., the child was not administered the item because he or she did not answer enough items correctly to advance to this item). The "Not administered" code is different than a system missing code in that only those children who were administered the Numbers Reversed subtask could have a "Not administered" code. If a child was not administered the Numbers Reversed subtask at all, that case would have a missing code for the Numbers Reversed scores. Variable names for the item-level data from the fall kindergarten assessments begin with "C1," and variable names for the item-level data from the spring kindergarten assessments begin with "C2." Variable descriptions for these items indicate the length of the digit sequence (e.g., C1 Numbers Reversed Two-digit sequence #1). Numbers Reversed was administered in Spanish for children routed through the assessment in Spanish. Data from English and Spanish administrations are combined into the same item-level variables.

In addition to the item-level data, three scores developed using guidelines from the publisher scoring materials are included in the data file for Numbers Reversed. Before analyzing the Numbers Reversed data, it is important that researchers understand the characteristics of these scores and how these characteristics may affect the analysis and interpretation of the Numbers Reversed data in the context of the ECLS-K:2011.

The three scores developed using publisher guidelines are a *W* score, a standard score, and percentile rank. Depending on the research question and analysis being conducted, one of the scores may be more preferable than another. For example, the *W* score may be best for a longitudinal analysis, whereas the percentile rank and standardized score may be better suited for an analysis focusing on one point in time. The descriptions below provide more information about which score may be better suited for a given analysis.[1]

The *W* score, a type of standardized score, is a special transformation of the Rasch ability scale and provides a common scale of equal intervals that represents both a child's ability and the task

[1] More information on these publisher scores can be found in the *Woodcock-Johnson III Test of Achievement Examiner's Manual: Standard and Extended Batteries* (Mather and Woodcock 2001).

difficulty. The *W* scale is particularly useful for the measurement of growth and can be considered a growth scale. Typically, the *W* scale has a mean of 500 and standard deviation of 100. Furthermore, the publisher of the Woodcock-Johnson III (WJ III) has set the mean to the average of performance for a child of 10 years, 0 months. This means that it would be expected that most children younger than 10 years, 0 months would obtain *W* scores lower than the mean, and most older children would be expected to have scores above the mean. Also, as children develop with age, it would be expected that the child's *W* score would increase to reflect growth. For example, when a child's *W*-ability score increases from 420 to 440, this indicates growth, and this would be the same amount of growth in the measured ability as any other student who gained 20 *W* points elsewhere on the measurement scale.

As mentioned above, the *W* score is an equal-interval scale, suited for analyses such as correlations and regressions. Higher *W* scores indicate that a child provided more correct responses and generally indicate that a child was able to correctly respond to at least some longer-number sequences. The *W* score accounts for only the total number of administered sequences answered correctly and does not reflect the pattern of responses, meaning that the *W* score does not indicate how many of each length number sequence the child answered correctly. As noted above, the data file includes item-level data that can be used to examine patterns of response.

The *W* score for each child in the ECLS-K:2011 was determined using norming data provided by the publisher. More specifically, a sample child was assigned the *W* score from the publisher norming data that was associated with the child's raw number-right score, the child's age (in months), and the language of administration. Norming data were provided separately for English and Spanish administrations of the task. Publisher materials indicate that the *W* scores earned on English administrations of the Numbers Reversed task are comparable to *W* scores earned on Spanish administrations of the task; however, differences related to precision of measurement in the norming samples result in different *W* scores for the same raw-number right score depending on the language of administration. For example, the lowest earnable *W* score on the English administration of the Numbers Reversed task is 403 (equivalent to a raw score of 0), and the lowest earnable *W* score on the Spanish administration is 393 (equivalent to raw score of 0). While this difference in the *W* scores between English and Spanish administration is largest at the lower end of the *W* distribution, the difference occurs along the entirety of the *W* distribution. For example, a raw score of 11 corresponds to a *W* score of 496 in the English administration norming data and a *W* score of 494 in the Spanish administration norming data. The data file includes one *W* score variable per round of data collection that contains data for all children administered the Numbers Reversed task, regardless of the language of administration.

Researchers who want to account for language of administration in their analyses can use the variables X1FLSCRN and X2FLSCRN, which are also in the data file, to identify which cases were administered Numbers Reversed in English and which cases were administered Numbers Reversed in Spanish.

Although the *W* score is reflective of the average performance of 10-year-olds, and the ECLS-K:2011 children were in kindergarten in the base-year collection, it is included in the data file because it sets a baseline that can be used to measure changes in children's working memory longitudinally across all rounds of the study. Also, it will facilitate comparisons of the ECLS-K:2011 data with data from other studies that include the Numbers Reversed task. Users should keep in mind that most ECLS-K:2011 sample children were 5 or 6 years old during the kindergarten data collections and that the *W* scores compare their performance to that of 10-year-olds. As a result, *W* scores from the ECLS-K:2011 sample appear to show that the ECLS-K:2011 children demonstrated below average performance on this task.

A score of 403 (393 for Spanish) is potentially a meaningful baseline value for the ability level of children who are unable to answer any items correctly. Over time, as children develop more ability that is measureable by the WJ III Numbers Reversed task, the study will be able to compare their baseline Numbers Reversed *W* score (either fall kindergarten or spring kindergarten) with their scores across future administrations of the task. However, researchers should understand that a score of 0 is an imprecise measure of children's ability in the area of working memory, because it is unknown how close a child was to answering at least one item correctly.

In fall kindergarten approximately 40 percent of students did not demonstrate sufficient skills as measured by this assessment to score above the lowest scalable score (403 for English assessment and 393 for Spanish assessment). In spring kindergarten, approximately 20 percent of students did not score above the lowest scalable score (403 for English, 393 for Spanish). Another factor that may contribute to the large number of children scoring 403 (and 393 for Spanish) is that some ECLS-K:2011 assessors did not properly administer the practice items, which may have resulted in some children never fully understanding what they were being asked to do during the Numbers Reversed task. During field observations of the assessors, it was noted that when children did not correctly answer the first practice item, there were inconsistencies in the administration of additional practice items. It is not possible to determine the extent to which improper administration of the practice items affected the results. However, researchers should keep in mind that this may have affected performance for some,

though not all, children. Researchers need to decide how to handle the 403 (393 for Spanish) scores in their analyses; the decision for how to do so is left up to the analyst based on his or her analytic goals.

Both the *standard score* and the *percentile score*, which indicate children's status relative to their peers, are age-normed transformations of the data. That is, both of these scores are relative to *same-aged* subjects in the WJ III norming sample. The standard score created by the publisher has a mean of 100 and a standard deviation of 15. The score is a linear transformation of a *z* score (mean of 0 and a standard deviation of 1), which is derived from a person's achieved *W* score. The percentile rank describes performance on a scale from 1 to 99 relative to the performance of subjects in the WJ III norming sample that are at the same age as the ECLS-K:2011 subjects.

Like the *W* score, the standard scores and the percentile scores in the data file contain data from both the English and Spanish administrations of the Numbers Reversed task. Standard scores and percentile scores are a function of the child's age at assessment. The publisher's scoring protocols result in standard and percentile scores that extend to slightly lower ages for children who were administered the task in Spanish compared to children who were administered the task in English, again due to differences in the precision of measurement within the norming samples. Children 62 months and younger who were administered the Numbers Reversed task in English, and who earned a raw score of 0 or 1, have a *W* score but do not have a standard score or percentile score (*W* scores are a function of the number correct and not a function of age). However, all children who were administered this task in Spanish, including those aged 62 months and younger, have a *W* score, a standard score, and a percentile score, regardless of their raw score. Again, there are variables in the data file indicating language of administration (X1FLSCRN and X2FLSCRN) that analysts may want to include in their analytic models.

Standard scores and percentile ranks lend themselves to different interpretations. Standard scores and percentile ranks are *not* essentially the same. Standard scores are deviation-based scores, based upon a mean and standard deviation that remain constant across the entire range. They are interval data, where values are separated by a constant interval that maintains the same meaning across the full range. Percentile ranks are neither interval data nor constant and cannot be used interchangeably with standardized scores. As such, standard scores are most appropriately used for comparisons across children and between groups; *W* scores (also a deviation-based score metric) are most appropriately used to look at growth over time, where age-normed standard scores may remain relatively constant with an age-expected rate of growth. Percentiles are less ideal for longitudinal analyses; although they can be used to

examine relative rank order consistency across time periods, the *W* scores would be better to assess change and/or stability across time.

6.2.1 Mean Scores

The variable names, descriptions, value ranges, weighted means, and standard deviations for the fall kindergarten and spring kindergarten Numbers Reversed scores are shown in table 6-4. In looking at the weighted means, researchers should keep in mind that the *W* score, the standard score, and the percentile score are age-normed scores, with the *W* score normed to the average 10–year-old and the standard and percentile scores normed to same-age peers in the WJ III norming sample. The low mean for the *W* score in the ECLS-K:2011 may be attributed to the derivation of the score being a comparison to the average 10-year-old or to differences between the ECLS-K:2011 population and the WJ III norming sample.[2] The standard score and the percentile rank also show a lower mean in the ECLS-K:2011, which may also be attributable to differences between the ECLS-K:2011 population and the WJ III norming sample.

Table 6-4. Numbers Reversed variable names, descriptions, value ranges, weighted means, and standard deviations: School year 2010–11

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1NRWABL	X1 Numbers Reversed *W*-ability score	15,600	393-581	432.56	30.028
X1NRSSCR	X1 Numbers Reversed standard score	14,440	45-175	93.10	16.510
X1NRPERC	X1 Numbers Reversed percentile rank	14,440	0-100	37.89	31.787
X2NRWABL	X2 Numbers Reversed *W*-ability score	17,150	393-572	449.49	30.412
X2NRSSCR	X2 Numbers Reversed standard score	17,120	40-175	94.92	17.017
X2NRPERC	X2 Numbers Reversed percentile rank	17,120	0-100	42.44	30.970

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Means and standard deviations for the Numbers Reversed scores are provided by data collection round and by child characteristics in tables 6-5 through 6-7.

[2] Normative data for the WJ III were gathered from 8,818 subjects in more than 100 geographically diverse U.S. communities. The kindergarten through 12th grade sample was composed of 4,783 subjects. The norming sample was selected to be representative of the U.S. population from age 24 months to age 90 years and older. Subjects were randomly selected within a stratified sampling design that controlled for the following 10 specific community and subject variables: census region (Northeast, Midwest, South, West); community size (city and urban, larger community, smaller community, rural area); sex; race (White, Black, American Indian, Asian and Pacific Islander); Hispanic or non-Hispanic; type of school (elementary, secondary, public, private, home); type of college/university (2-year, 4-year, public, private); education of adults; occupational status of adults; occupation of adults in the labor force.

Table 6-5. Mean Numbers Reversed *W*-ability score, by data collection round and child characteristics: School year 2010–11

	Fall 2010			Spring 2011		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	15,600	432.56	30.028	17,150	449.49	30.412
Sex						
Male	7,920	431.37	29.959	8,710	448.03	31.363
Female	7,610	433.92	30.061	8,370	451.09	29.298
Race/ethnicity						
White, non-Hispanic	7,590	439.47	29.605	8,090	456.03	28.112
Black, non-Hispanic	2,110	423.12	27.209	2,240	439.89	30.710
Hispanic	3,760	421.37	27.096	4,300	439.29	30.871
Asian, non-Hispanic	1,140	439.63	31.873	1,430	455.87	30.639
Hawaiian, Other Pacific Islander, non-Hispanic	80	429.62	29.066	110	451.42	29.424
American Indian/Alaska Native, non-Hispanic	150	430.15	29.814	160	447.47	30.772
More than one race, non-Hispanic	700	438.34	30.348	750	454.92	28.621
School type						
Public school	13,520	431.43	29.842	15,000	448.37	30.569
Private school	2,080	441.70	29.958	2,130	458.60	27.472

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 6-6. Mean Numbers Reversed standard score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	14,440	93.10	16.510	17,120	94.92	17.017
Sex						
Male	7,350	92.07	16.540	8,700	93.86	17.509
Female	7,040	94.24	16.410	8,360	96.04	16.421
Race/ethnicity						
White, non-Hispanic	7,200	96.23	16.161	8,090	98.01	16.155
Black, non-Hispanic	1,920	87.83	15.644	2,230	89.54	17.143
Hispanic	3,370	87.61	15.200	4,290	90.07	16.840
Asian, non-Hispanic	1,030	99.53	17.680	1,420	100.41	17.401
Hawaiian, Other Pacific Islander, non-Hispanic	80	91.56	15.812	110	96.20	16.833
American Indian/Alaska Native, non-Hispanic	140	91.54	16.007	160	93.18	16.670
More than one race, non-Hispanic	660	96.04	16.684	750	97.65	16.337
School type						
Public school	12,560	92.31	16.380	14,990	94.21	17.050
Private school	1,880	99.69	16.113	2,120	100.69	15.565

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 6-7. Mean Numbers Reversed percentile rank, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	14,440	37.89	31.787	17,120	42.44	30.970
Sex						
Male	7,350	35.99	31.609	8,700	40.78	31.428
Female	7,040	39.99	31.852	8,360	44.23	30.396
Race/ethnicity						
White, non-Hispanic	7,200	43.80	31.502	8,090	48.09	29.979
Black, non-Hispanic	1,920	28.06	29.424	2,230	33.08	29.952
Hispanic	3,370	27.39	28.911	4,290	33.41	29.976
Asian, non-Hispanic	1,030	50.19	33.672	1,420	52.15	31.994
Hawaiian, Other Pacific Islander, non-Hispanic	80	34.80	30.328	110	44.16	29.644
American Indian/Alaska Native, non-Hispanic	140	35.15	31.654	160	39.67	30.394
More than one race, non-Hispanic	660	43.43	32.148	750	47.28	30.465
School type						
Public school	12,560	36.36	31.447	14,990	41.16	30.872
Private school	1,880	50.54	31.773	2,120	52.84	29.757

NOTE: Estimates weighted by W1C0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

This page intentionally left blank.

7. PSYCHOMETRIC CHARACTERISTICS OF THE INDIRECT MEASURES

This chapter describes the indirect measures included in the fall and spring kindergarten rounds of data collection. Parents and teachers of sampled children completed the indirect measures assessing children's social skills. Teachers also completed indirect measures of children's executive functioning skills.

7.1 Teacher Measures

7.1.1 *Children's Behavior Questionnaire (CBQ)*

The fall kindergarten and spring kindergarten child-level teacher questionnaires included 12 items from the *Short Form of the Children's Behavior Questionnaire* (Putnam and Rothbart 2006)[1] asking teachers to indicate how often the ECLS-K:2011 children in their classroom exhibited certain social skills and behaviors related to inhibitory control and attentional focusing, two indicators of executive functioning. Teachers were presented with statements about how the children might have reacted to a number of situations in the past 6 months and were asked to indicate how "true" or "untrue" those statements were about that child on a 7-point scale ranging from extremely untrue to extremely true, with a middle option of "neither true nor untrue." If a statement or situation did not apply to that child, the teacher could indicate "not applicable."

The data file includes two scale scores derived from the CBQ items: (1) attentional focus and (2) inhibitory control. The scale scores were developed based on guidelines from the publisher and included all six items from the Attentional Focusing subscale and all six items from the Inhibitory Control subscale from the CBQ *Short Form*. The score on each scale is the mean rating on the items included in the scale. A score was computed when the respondent provided a rating on at least four of the six items that composed the scale. Higher scale scores on the attentional focus scale indicate that the child exhibited more behaviors that demonstrate the ability to focus attention on cues in the environment that are relevant to the task in hand. Higher scale scores on the inhibitory control scale indicate that the child exhibited more behaviors that demonstrate the ability to resist a strong inclination to do one thing and instead to do what is most appropriate or needed. The variable names, descriptions, value ranges, weighted means, and

[1]The *Children's Behavior Questionnaire* is a copyrighted instrument and has been used with permission.

standard deviations for these scales are shown in table 7-1. The attentional focus scale has an internal consistency reliability coefficient (Cronbach's alpha) of .87 for each round of data collection. The inhibitory control scale also has a Cronbach's alpha of .87 for each round of data collection. Data for the individual *Children's Behavior Questionnaire* items are not included in the data file because of copyright restrictions.

Table 7-1. *Children's Behavior Questionnaire* variable names, descriptions, value ranges, weighted means, and standard deviations: School year 2010–11

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1ATTNFS	X1 Teacher Report Attentional Focus	14,560	1–7	4.68	1.325
X1INBCNT	X1 Teacher Report Inhibitory Control	14,560	1–7	4.88	1.291
X2ATTNFS	X2 Teacher Report Attentional Focus	15,940	1–7	4.90	1.328
X2INBCNT	X2 Teacher Report Inhibitory Control	15,930	1–7	5.06	1.288

NOTE: Fall kindergarten estimates (X1) weighted by W1T0. Spring kindergarten estimates (X2) weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to these scales come from the *Children's Behavior Questionnaire* (Putnam and Rothbart 2006).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Mean scores for the *Children's Behavior Questionnaire* scales are presented by data collection round and by child characteristics in tables 7-2 and 7-3.

Table 7-2. Mean *Children's Behavior Questionnaire* attentional focus score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	14,560	4.68	1.325	15,940	4.90	1.328
Sex						
Male	7,440	4.40	1.345	8,110	4.63	1.351
Female	7,060	4.98	1.237	7,770	5.19	1.240
Race/ethnicity						
White, non-Hispanic	7,170	4.77	1.310	7,710	4.99	1.320
Black, non-Hispanic	1,970	4.41	1.362	2,090	4.61	1.370
Hispanic	3,460	4.60	1.307	3,870	4.83	1.306
Asian, non-Hispanic	1,030	4.97	1.234	1,260	5.28	1.195
Hawaiian, Other Pacific Islander, non-Hispanic	70	4.67	1.379	100	4.82	1.403
American Indian/Alaska Native, non- Hispanic	150	4.68	1.409	150	4.78	1.339
More than one race, non-Hispanic	660	4.66	1.369	700	4.89	1.314
School type						
Public school	12,630	4.64	1.331	13,900	4.87	1.337
Private school	1,940	4.99	1.225	2,030	5.16	1.223

NOTE: Fall kindergarten estimates weighted by W1T0. Spring kindergarten estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to these scales come from the *Children's Behavior Questionnaire* (Putnam and Rothbart 2006). The range of possible values is 1 to 7. Detail may not sum to total due to rounding and/or missing data.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 7-3. Mean *Children's Behavior Questionnaire* inhibitory control score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	14,560	4.88	1.291	15,930	5.06	1.288
Sex						
Male	7,430	4.59	1.325	8,100	4.76	1.330
Female	7,070	5.19	1.177	7,770	5.38	1.160
Race/ethnicity						
White, non-Hispanic	7,180	4.95	1.277	7,700	5.13	1.288
Black, non-Hispanic	1,970	4.60	1.356	2,090	4.72	1.367
Hispanic	3,440	4.87	1.254	3,870	5.07	1.235
Asian, non-Hispanic	1,020	5.07	1.223	1,260	5.30	1.147
Hawaiian, Other Pacific Islander, non-Hispanic	70	4.96	1.407	100	4.99	1.340
American Indian/Alaska Native, non-Hispanic	150	4.84	1.312	150	5.06	1.212
More than one race, non-Hispanic	660	4.81	1.373	710	5.04	1.271
School type						
Public school	12,620	4.86	1.295	13,910	5.04	1.291
Private school	1,940	5.07	1.241	2,020	5.17	1.255

NOTE: Fall kindergarten estimates weighted by W1T0. Spring kindergarten estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to these scales come from the *Children's Behavior Questionnaire* (Putnam and Rothbart 2006). The range of possible values is 1 to 7. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

7.1.2 Teacher-Reported Social Skills

In both the fall and spring kindergarten collections, teachers reported how often the ECLS-K:2011 children in their classroom exhibited certain social skills and behaviors using a four-option frequency scale ranging from "Never" to "Very Often." Teachers also had the option of indicating that they had not had an opportunity to observe the described behavior for the child being asked about. The items tapping children's social skills and behaviors are based on items from the *Social Skills Rating System* (SSRS)[2] and included in the self-administered child-level teacher questionnaire. The social skills battery includes some items taken verbatim from the SRSS, some items that are modifications of original

[2]The *Social Skills Rating System* is a copyrighted instrument (1990 NCS Pearson) and has been adapted with permission.

SRSS items, and some items that measure the same kinds of skills and behaviors captured in the SRSS but use wording developed specifically for the ECLS studies.

Four social skills scales were developed based on teachers' responses to these questionnaire items. The scores were derived in the same way as those reported for the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K) to enable comparisons between the two studies. The score on each scale is the mean rating on the items included in the scale. The four teacher scales are as follows: self-control (4 items), interpersonal skills (5 items), externalizing problem behaviors (5 items), and internalizing problem behaviors (4 items). A score was computed when the respondent provided a rating on at least a minimum number of the items that composed the scale. The minimum number of items that were required to compute a score were as follows: self-control (3 out of 4 items), interpersonal skills (4 out of 5 items), externalizing problem behaviors (4 out of 5 items), and internalizing problem behaviors (3 out of 4 items). Higher scores indicate that the child exhibited the behavior represented by the scale more often (e.g., higher self-control scores indicate that the child exhibited behaviors indicative of self-control more often; higher externalizing problem behavior scores indicate that the child exhibited more externalizing problem behaviors more often). Variable names for the fall kindergarten and spring kindergarten teacher scale scores, descriptions, value ranges, weighted means, and standard deviations for these scales are shown in table 7-4. Data for the individual items contributing to each scale are not included in the data file because of copyright restrictions.

Table 7-4. Teacher-reported social skills scales variable names, descriptions, value ranges, weighted means, and standard deviations: School year 2010–11

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1TCHCON	X1 Teacher Report Self-Control	13,550	1–4	3.07	0.629
X1TCHPER	X1 Teacher Report Interpersonal Skills	13,710	1–4	2.98	0.640
X1TCHEXT	X1 Teacher Report Externalizing Problem Behaviors	14,390	1–4	1.61	0.632
X1TCHINT	X1 Teacher Report Internalizing Problem Behaviors	14,240	1–4	1.47	0.494
X2TCHCON	X2 Teacher Report Self-Control	15,800	1–4	3.18	0.635
X2TCHPER	X2 Teacher Report Interpersonal Skills	15,800	1–4	3.14	0.649
X2TCHEXT	X2 Teacher Report Externalizing Problem Behaviors	15,900	1–4	1.64	0.637
X2TCHINT	X2 Teacher Report Internalizing Problem Behaviors	15,870	1–4	1.51	0.500

NOTE: Fall kindergarten estimates (X1) weighted by W1T0. Spring kindergarten estimates (X2) weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 7-5 presents the internal consistency reliability coefficients (Cronbach's alpha) for the self-control, interpersonal skills, externalizing problem behaviors, and internalizing problem behaviors scales derived from information reported by the teacher.

Table 7-5. Reliability estimates for the teacher-reported social skills scales: School year 2010–11

Variable name	Description	Number of items	Reliability coefficient
X1TCHCON	X1 Teacher Report Self-Control	4	.81
X1TCHPER	X1 Teacher Report Interpersonal Skills	5	.86
X1TCHEXT	X1 Teacher Report Externalizing Problem Behaviors	5	.88
X1TCHINT	X1 Teacher Report Internalizing Problem Behaviors	4	.79
X2TCHCON	X2 Teacher Report Self-Control	4	.82
X2TCHPER	X2 Teacher Report Interpersonal Skills	5	.87
X2TCHEXT	X2 Teacher Report Externalizing Problem Behaviors	5	.89
X2TCHINT	X2 Teacher Report Internalizing Problem Behaviors	4	.78

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

To further explore the factor structure of the teacher-reported social skills scales, principal components factor analyses with varimax rotation were conducted for the social skills items in both fall and spring kindergarten. The factor analyses with the teacher-reported social skills items specified the extraction of three factors in both fall and spring kindergarten. The eigenvalues and proportion of variance accounted for by each component are listed in tables 7-6 and 7-7. The three factors account for a total of 63.8 percent of the variance in fall kindergarten and 64.7 percent of the variance in spring kindergarten. The factor structure for externalizing problem behaviors and internalizing problem behaviors generally matched the planned composite structure for both fall and spring kindergarten. Self-control and interpersonal skills, on the other hand, loaded onto the same factor and appear to be measuring the same construct. However, because the internal consistency of the self-control and interpersonal skills scales supported the creation of separate scores, separate scores for self-control and interpersonal skills were created in the same way as the variables were for the ECLS-K to make the scores directly comparable.

Table 7-6. Eigenvalues and proportion of variance accounted for by the three factors extracted in principal components factor analysis with fall kindergarten teacher-reported social skills data: Fall 2010

Result	Factor 1: Self-control/ Interpersonal	Factor 2: Externalizing	Factor 3: Internalizing
Eigenvalue	7.87	2.15	1.46
Proportion of variance accounted for by component	.44	.12	.08.

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010.

Table 7-7. Eigenvalues and proportion of variance accounted for by the three factors extracted in principal components factor analysis with spring kindergarten teacher-reported social skills data: Spring 2011

Result	Factor 1: Self-control/ Interpersonal	Factor 2: Externalizing	Factor 3: Internalizing
Eigenvalue	8.22	2.01	1.35
Proportion of variance accounted for by component	.46	.11	.08

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2011.

Mean scores for the teacher-reported social skills subscales are presented by data collection round and child characteristics in tables 7-8 through 7-11.

Table 7-8. Mean teacher-reported self-control score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	13,550	3.07	0.629	15,800	3.18	0.635
Sex						
Male	6,930	2.96	0.643	8,030	3.07	0.649
Female	6,570	3.19	0.591	7,710	3.29	0.598
Race/ethnicity						
White, non-Hispanic	6,750	3.11	0.613	7,680	3.22	0.620
Black, non-Hispanic	1,820	2.94	0.677	2,060	2.99	0.692
Hispanic	3,160	3.06	0.624	3,820	3.18	0.620
Asian, non-Hispanic	940	3.11	0.612	1,240	3.25	0.588
Hawaiian, Other Pacific Islander, non-Hispanic	70	3.05	0.702	100	3.22	0.647
American Indian/Alaska Native, non-Hispanic	140	3.13	0.655	150	3.25	0.611
More than one race, non-Hispanic	610	3.05	0.621	700	3.13	0.654
School type						
Public school	11,700	3.06	0.632	13,760	3.17	0.636
Private school	1,850	3.12	0.597	2,040	3.19	0.622

NOTE: Fall kindergarten estimates weighted by W1T0. Spring kindergarten estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 7-9. Mean teacher-reported interpersonal skills score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	13,710	2.98	0.640	15,800	3.14	0.649
Sex						
Male	6,940	2.87	0.644	8,020	3.01	0.652
Female	6,720	3.10	0.613	7,720	3.27	0.618
Race/ethnicity						
White, non-Hispanic	6,860	3.02	0.624	7,680	3.18	0.639
Black, non-Hispanic	1,840	2.88	0.672	2,060	2.98	0.677
Hispanic	3,180	2.95	0.643	3,810	3.14	0.640
Asian, non-Hispanic	940	2.96	0.640	1,240	3.12	0.632
Hawaiian, Other Pacific Islander, non-Hispanic	60	2.99	0.648	100	3.16	0.703
American Indian/Alaska Native, non-Hispanic	140	2.97	0.680	150	3.16	0.673
More than one race, non-Hispanic	630	2.98	0.644	710	3.12	0.661
School type						
Public school	11,850	2.96	0.642	13,780	3.13	0.652
Private school	1,860	3.11	0.607	2,020	3.19	0.624

NOTE: Fall kindergarten estimates weighted by W1T0. Spring kindergarten estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 7-10. Mean teacher-reported externalizing problem behaviors score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	14,390	1.61	0.632	15,900	1.64	0.637
Sex						
Male	7,340	1.75	0.677	8,100	1.76	0.678
Female	6,990	1.47	0.545	7,750	1.51	0.563
Race/ethnicity						
White, non-Hispanic	7,110	1.60	0.620	7,710	1.62	0.625
Black, non-Hispanic	1,940	1.76	0.708	2,070	1.85	0.744
Hispanic	3,410	1.57	0.599	3,850	1.59	0.596
Asian, non-Hispanic	1,000	1.48	0.527	1,260	1.48	0.508
Hawaiian, Other Pacific Islander, non-Hispanic	70	1.58	0.667	100	1.59	0.691
American Indian/Alaska Native, non-Hispanic	140	1.63	0.622	150	1.66	0.573
More than one race, non-Hispanic	650	1.69	0.699	700	1.69	0.645
School type						
Public school	12,470	1.62	0.635	13,870	1.64	0.640
Private school	1,910	1.59	0.604	2,030	1.64	0.616

NOTE: Fall kindergarten estimates weighted by W1T0. Spring kindergarten estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 7-11. Mean teacher-reported internalizing problem behaviors score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	14,240	1.47	0.494	15,870	1.51	0.500
Sex						
Male	7,260	1.49	0.505	8,070	1.53	0.512
Female	6,930	1.44	0.482	7,740	1.50	0.485
Race/ethnicity						
White, non-Hispanic	7,080	1.47	0.477	7,710	1.51	0.484
Black, non-Hispanic	1,900	1.46	0.509	2,060	1.54	0.530
Hispanic	3,360	1.46	0.516	3,850	1.51	0.510
Asian, non-Hispanic	980	1.40	0.461	1,240	1.40	0.411
Hawaiian, Other Pacific Islander, non-Hispanic	70	1.31	0.369	100	1.39	0.399
American Indian/Alaska Native, non-Hispanic	140	1.53	0.513	140	1.64	0.607
More than one race, non-Hispanic	640	1.52	0.541	710	1.57	0.553
School type						
Public school	12,340	1.47	0.496	13,830	1.52	0.502
Private school	1,900	1.44	0.478	2,040	1.48	0.478

NOTE: Fall kindergarten estimates weighted by W1T0. Spring kindergarten estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

7.1.3 Teacher-Reported Approaches to Learning Items and Scale

The child-level teacher questionnaire included seven items, referred to as "Approaches to Learning" items, that asked the teachers to report how often the ECLS-K:2011 children in their classroom exhibited a selected set of learning behaviors (keeps belongings organized; shows eagerness to learn new things; works independently; easily adapts to changes in routine; persists in completing tasks; pays attention well; and follows classroom rules).[3] These items were presented in the same set of items as the

[3]The Approaches to Learning teacher items were developed specifically for the ECLS-K; they are not taken from an existing source. These are the same items that were fielded as part of what was called the Teacher Social Rating Scale in the ECLS-K. The first six items (i.e., keeps belongings organized; shows eagerness to learn new things; works independently; easily adapts to changes in routine; persists in completing tasks; pays attention well) were included in the Teacher Social Rating Scale of the kindergarten round in ECLS-K. The seventh item (i.e., follows classroom rules) was added in the first-grade round of ECLS-K.

social skills items based on the *Social Skills Rating System* (described in section 7.1.2), and teachers used the same frequency scale to report how often each child demonstrated the behaviors described. The Approaches to Learning Scale score is the mean rating on the seven items included in the scale. A score was computed when the respondent provided a rating on at least four of the seven items that composed the scale. Higher scale scores indicate that the child exhibited positive learning behaviors more often. The variable names, descriptions, value ranges, weighted means, and standard deviations for the fall kindergarten and spring kindergarten teacher Approaches to Learning Scale scores are shown in table 7-12. The Approaches to Learning Scale has an internal consistency reliability estimate of .91 for each round of data collection, as measured by Cronbach's alpha. Additionally, the item-level data for the teacher-reported Approaches to Learning items are included in the data file along with the other child-level teacher questionnaire data. Variable names for the item-level data from the fall kindergarten child-level teacher questionnaire begin with "T1," and variable names for the spring kindergarten child-level teacher questionnaire begin with "T2."

Table 7-12. Teacher-reported Approaches to Learning Scale variable names, descriptions, value ranges, weighted means, and standard deviations: School year 2010–11

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1TCHAPP	X1 Teacher Report Approaches to Learning	14,770	1–4	2.93	0.681
X2TCHAPP	X2 Teacher Report Approaches to Learning	15,980	1–4	3.09	0.688

NOTE: Fall kindergarten estimates (X1) weighted by W1T0. Spring kindergarten estimates (X2) weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Mean scores for the teacher-reported Approaches to Learning Scale are presented by data collection round and child characteristics in table 7-13.

Table 7-13. Mean teacher-reported Approaches to Learning Scale scores, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	14,770	2.93	0.681	15,980	3.09	0.688
Sex						
Male	7,540	2.78	0.680	8,130	2.93	0.695
Female	7,170	3.08	0.646	7,790	3.26	0.639
Race/ethnicity						
White, non-Hispanic	7,280	2.97	0.669	7,730	3.13	0.677
Black, non-Hispanic	2,000	2.80	0.702	2,090	2.93	0.717
Hispanic	3,500	2.88	0.677	3,890	3.07	0.684
Asian, non-Hispanic	1,040	3.04	0.685	1,260	3.21	0.662
Hawaiian, Other Pacific Islander, non-Hispanic	80	2.89	0.680	100	3.05	0.775
American Indian/Alaska Native, non-Hispanic	150	2.98	0.729	150	3.09	0.668
More than one race, non-Hispanic	670	2.91	0.693	710	3.07	0.690
School type						
Public school	12,790	2.91	0.683	13,940	3.08	0.693
Private school	1,980	3.07	0.649	2,040	3.16	0.646

NOTE: Fall kindergarten estimates weighted by W1T0. Spring kindergarten estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

7.1.4 *Student-Teacher Relationship Scale*

The *Student-Teacher Relationship Scale* (Pianta and Steinberg 2001) is a 15-item, teacher-reported measure of closeness and conflict between the teacher and child. As part of the spring kindergarten child-level teacher questionnaire, the teacher was presented with 15 descriptive statements about his or her relationship with the ECLS-K:2011 child and asked to indicate the degree to which each statement applied to their relationship using a 5-point scale ranging from "definitely does not apply" to "definitely applies." Two scales were developed based on guidelines from the publisher: closeness and conflict. The closeness scale score is the average rating on the seven items included in the scale, while the conflict scale score is the average rating on the eight items included in that scale. A score was computed when the respondent provided a rating on at least five of the seven or eight items that composed the

scales. The closeness subscale is a measure of the affection, warmth, and open communication that the teacher experiences with the student. The conflict subscale is a measure of the teacher's perception of the negative and conflicting aspects of the teacher's relationship with the student. High scale scores on the closeness scale indicate that the teacher perceived he or she had a closer relationship with the child. High scale scores on the conflict scale indicate that the teacher perceived his or her relationship with the child to be characterized by more conflict. The variable names, descriptions, value ranges, weighted means, and standard deviations for these scales are shown in table 7-14. The conflict scale has an internal consistency reliability estimate of .89, and the closeness scale also has a reliability estimate of .89, as measured by Cronbach's alpha. Data for the individual *Student-Teacher Relationship Scale* items are not included in the data file because of copyright restrictions.

Table 7-14. *Student-Teacher Relationship Scale* variable names, descriptions, value ranges, weighted means, and standard deviations: Spring 2011

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation[1]
X2CLSNSS	X2 Teacher Report Closeness	15,960	1–5	4.36	0.635
X2CNFLCT	X2 Teacher Report Conflict	15,960	1–5	1.63	0.800

NOTE: Estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Steinberg 1992).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2011.

Mean scores for the *Student-Teacher Relationship Scale* are presented by child characteristics in tables 7-15 and 7-16. As noted earlier, this scale was included only in the spring 2011 collection, so scores are not presented by round of data collection.

Table 7-15. Mean *Student-Teacher Relationship Scale* teacher-reported closeness score, by child characteristics: Spring 2011

Characteristic	Spring 2011		
	Number	Mean	Standard deviation
Total sample	15,960	4.36	0.635
Sex			
Male	8,130	4.26	0.658
Female	7,780	4.47	0.589
Race/ethnicity			
White, non-Hispanic	7,720	4.43	0.592
Black, non-Hispanic	2,080	4.28	0.678
Hispanic	3,890	4.28	0.667
Asian, non-Hispanic	1,260	4.26	0.692
Hawaiian, Other Pacific Islander, non-Hispanic	100	4.29	0.677
American Indian/Alaska Native, non-Hispanic	150	4.33	0.635
More than one race, non-Hispanic	710	4.35	0.622
School type			
Public school	13,930	4.34	0.639
Private school	2,030	4.50	0.582

NOTE: Estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Steinberg 1992). The range of possible values is 1 to 5. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2011.

Table 7-16. Mean *Student-Teacher Relationship Scale* teacher-reported conflict score, by child characteristics: Spring 2011

Characteristic	Spring 2011		
	Number	Mean	Standard deviation
Total sample	15,960	1.63	0.800
Sex			
Male	8,130	1.77	0.863
Female	7,780	1.49	0.701
Race/ethnicity			
White, non-Hispanic	7,720	1.60	0.795
Black, non-Hispanic	2,080	1.88	0.948
Hispanic	3,890	1.57	0.710
Asian, non-Hispanic	1,260	1.49	0.634
Hawaiian, Other Pacific Islander, non-Hispanic	100	1.65	0.847
American Indian/Alaska Native, non-Hispanic	150	1.74	0.786
More than one race, non-Hispanic	710	1.71	0.838
School type			
Public school	13,930	1.64	0.801
Private school	2,030	1.61	0.793

NOTE: Estimates weighted by W12T0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Steinberg 1992). The range of possible values is 1 to 5. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2011.

7.2 Parent Measures

7.2.1 Parent-Reported Social Skills

In both the fall and spring kindergarten parent interviews, parents were asked to report how often their child exhibited certain social skills and behaviors using the same frequency scale described earlier for the teacher-reported social skills items. These parent items also are based on items from the *Social Skills Rating System*.

Four social skills scales were developed based on parents' responses to these interview questions. The score on each scale is the mean rating on the items included in the scale. The four social-skill parent scales are as follows: self-control (5 items), social interaction (3 items), sad/lonely (4 items),

and impulsive/overactive behaviors (2 items). A score was computed when the respondent provided a rating on at least a minimum number of the items that composed the scale. The minimum number of items that were required to compute a score were as follows: self-control (4 out of 5 items), social interaction (2 out of 3 item), sad/lonely (3 out of 4 items), and impulsive/overactive (2 out of 2 items). Higher scores indicate that the child exhibited the behavior represented by the scale more often (e.g., higher self-control scores indicate that the child exhibited behaviors indicative of self-control more often; higher scores on the social interaction scale indicate that the child interacted with others in a positive way more often). The variable names, descriptions, value ranges, weighted means, and standard deviations for the fall kindergarten and spring kindergarten parent scores are shown in table 7-17. Data for the individual items contributing to each scale are not included in the data file because of copyright restrictions.

Table 7-17. Parent-reported social skills scales variable names, descriptions, value ranges, weighted means, and standard deviations: School year 2010–11

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1PRNCON	X1 Parent Report Self-Control	13,210	1–4	2.88	0.525
X1PRNSOC	X1 Parent Report Social Interaction	13,230	1–4	3.44	0.564
X1PRNSAD	X1 Parent Report Sad/Lonely	13,210	1–4	1.49	0.379
X1PRNIMP	X1 Parent Report Impulsive/Overactive	13,130	1–4	2.05	0.679
X2PRNCON	X2 Parent Report Self-Control	13,250	1–4	2.95	0.508
X2PRNSOC	X2 Parent Report Social Interaction	13,270	1–4	3.44	0.549
X2PRNSAD	X2 Parent Report Sad/Lonely	13,230	1–4	1.47	0.382
X2PRNIMP	X2 Parent Report Impulsive/Overactive	13,150	1–4	1.92	0.683

NOTE: Fall kindergarten estimates weighted by W1P0. Spring kindergarten estimates weighted by W2P0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 7-18 presents the internal consistency reliability coefficients (Cronbach's alpha) of the self-control, social interaction, and sad/lonely scales derived from information reported by the parent. Reliability statistics are not reported for the impulsive/overactive scale; that scale is computed from only two parent-reported items, which is not enough to calculate an alpha reliability.

Table 7-18. Reliability estimates for the parent-reported social skills scales: School year 2010–11

Variable name	Description	Number of items	Reliability coefficient
X1PRNCON	X1 Parent Report Self-Control	5	.73
X1PRNSOC	X1 Parent Report Social Interaction	3	.68
X1PRNSAD	X1 Parent Report Sad/Lonely	4	.56
X2PRNCON	X2 Parent Report Self-Control	5	.72
X2PRNSOC	X2 Parent Report Social Interaction	3	.67
X2PRNSAD	X2 Parent Report Sad/Lonely	4	.58

NOTE: Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

To further explore the factor structure of the parent-reported social skills scales, principal components factor analyses with varimax rotation were conducted for the social skills items in both fall kindergarten and spring kindergarten. The eigenvalues and proportion of variance accounted for by each component are listed in tables 7-19 and 7-20. Three factors were extracted in both analyses. These factors accounted for a total of 47.3 percent of the variance in fall kindergarten and 47.3 percent of the variance in spring kindergarten. The factors that were extracted matched the planned composite structure which was based upon the way the scales were created for the ECLS-K, except that the two items measuring hyperactivity and impulsivity loaded on the self-control factor. Because these items were conceptually meaningful and they were used in a separate scale in the ECLS-K, a separate impulsivity/overactivity scale score was also created for both fall and spring kindergarten in the ECLS-K:2011.

Table 7-19. Eigenvalues and proportion of variance accounted for by the three factors extracted in principal components factor analysis with fall kindergarten parent-reported social skills data: Fall 2010

Result	Factor 1: Self-Control	Factor 2: Social Interaction	Factor 3: Sad/Lonely
Eigenvalue	3.45	1.81	1.36
Proportion of variance accounted for by component	.25	.13	.10

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010.

Table 7-20. Eigenvalues and proportion of variance accounted for by the three factors extracted in principal components factor analysis with spring kindergarten parent-reported social skills data: Spring 2011

Result	Factor 1: Self-Control	Factor 2: Social Interaction	Factor 3: Sad/Lonely
Eigenvalue	3.46	1.78	1.39
Proportion of variance accounted for by component	.25	.13	.10

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2011.

Mean scores for the parent-reported social skills subscales are presented by data collection round and child characteristics in tables 7-21 through 7-24.

Table 7-21. Mean parent-reported self-control score, by data collection round and child characteristics: School year 2010–11

	Fall 2010			Spring 2011		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	13,210	2.88	0.525	13,250	2.95	0.508
Sex						
Male	6,790	2.85	0.531	6,810	2.92	0.513
Female	6,420	2.92	0.517	6,450	2.99	0.499
Race/ethnicity						
White, non-Hispanic	6,820	2.90	0.490	6,710	2.97	0.477
Black, non-Hispanic	1,690	2.90	0.596	1,480	2.96	0.580
Hispanic	2,980	2.84	0.562	3,150	2.89	0.537
Asian, non-Hispanic	880	2.94	0.474	1,100	3.05	0.448
Hawaiian, Other Pacific Islander, non- Hispanic	60	2.94	0.474	70	2.99	0.456
American Indian/Alaska Native, non-Hispanic	110	2.93	0.461	100	2.98	0.461
More than one race, non-Hispanic	660	2.88	0.527	650	2.93	0.521
School type						
Public school	11,390	2.88	0.532	11,470	2.94	0.514
Private school	1,810	2.96	0.453	1,670	3.06	0.435

NOTE: Fall kindergarten estimates weighted by W1P0. Spring kindergarten estimates weighted by W2P0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 7-22. Mean parent-reported social interaction score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	13,230	3.44	0.564	13,270	3.44	0.549
Sex						
Male	6,800	3.39	0.578	6,820	3.40	0.560
Female	6,430	3.48	0.545	6,460	3.48	0.533
Race/ethnicity						
White, non-Hispanic	6,830	3.50	0.511	6,710	3.52	0.492
Black, non-Hispanic	1,700	3.47	0.564	1,480	3.45	0.564
Hispanic	2,980	3.29	0.622	3,160	3.29	0.606
Asian, non-Hispanic	880	3.28	0.624	1,110	3.28	0.614
Hawaiian, Other Pacific Islander, non-Hispanic	60	3.49	0.545	70	3.42	0.484
American Indian/Alaska Native, non-Hispanic	110	3.60	0.525	100	3.44	0.506
More than one race, non-Hispanic	660	3.45	0.575	650	3.48	0.530
School type						
Public school	11,420	3.42	0.572	11,490	3.43	0.557
Private school	1,820	3.54	0.477	1,680	3.55	0.462

NOTE: Fall kindergarten estimates weighted by W1P0. Spring kindergarten estimates weighted by W2P0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 7-23. Mean parent-reported sad/lonely score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	13,210	1.95	0.379	13,230	1.47	0.382
Sex						
Male	6,790	1.48	0.378	6,800	1.47	0.384
Female	6,420	1.49	0.379	6,430	1.47	0.380
Race/ethnicity						
White, non-Hispanic	6,830	1.48	0.348	6,700	1.47	0.350
Black, non-Hispanic	1,690	1.49	0.419	1,470	1.47	0.413
Hispanic	2,980	1.47	0.409	3,140	1.46	0.425
Asian, non-Hispanic	880	1.62	0.412	1,100	1.53	0.399
Hawaiian, Other Pacific Islander, non-Hispanic	6260	1.53	0.378	70	1.62	0.441
American Indian/Alaska Native, non-Hispanic	110	1.51	0.341	100	1.55	0.458
More than one race, non-Hispanic	660	1.51	0.374	650	1.48	0.352
School type						
Public school	11,390	1.49	0.384	11,440	1.47	0.387
Private school	1,820	1.47	0.334	1,670	1.45	0.329

NOTE: Fall kindergarten estimates weighted by W1P0. Spring kindergarten estimates weighted by W2P0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Table 7-24. Mean parent-reported impulsive/overactive score, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	13,130	2.05	0.679	13,150	1.92	0.683
Sex						
Male	6,760	2.13	0.691	6,760	2.00	0.700
Female	6,380	1.96	0.654	6,390	1.84	0.654
Race/ethnicity						
White, non-Hispanic	6,800	2.03	0.657	6,690	1.89	0.643
Black, non-Hispanic	1,680	2.12	0.716	1,460	2.04	0.757
Hispanic	2,960	2.04	0.700	3,120	1.91	0.718
Asian, non-Hispanic	860	1.98	0.641	1,070	1.86	0.641
Hawaiian, Other Pacific Islander, non-Hispanic	60	2.10	0.676	70	2.10	0.642
American Indian/Alaska Native, non-Hispanic	110	2.08	0.696	100	1.97	0.625
More than one race, non-Hispanic	660	2.12	0.703	650	1.99	0.735
School type						
Public school	11,330	2.06	0.686	11,380	1.93	0.692
Private school	1,810	1.94	0.604	1,670	1.82	0.600

NOTE: Fall kindergarten estimates weighted by W1P0. Spring kindergarten estimates weighted by W2P0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

7.2.2 Parent-Reported Approaches to Learning Items and Scale

The parent interview included six items, referred to as "Approaches to Learning" items, that asked parents to report how often their child exhibited a selected set of learning behaviors (keep working at something until finished; show interest in a variety of things; concentrate on a task and ignore distractions; help with chores; eager to learn new things; creative in work and play).[4] These items were asked within the same set of items as the social skills items based on the *Social Skills Rating System* (described in section 7.2.1) in section SSQ (Social Skills, Problem Behaviors, and Approaches to

[4]The Approaches to Learning parent items were developed specifically for the ECLS-K; they are not taken from an existing source. These are the same items that were fielded as part of what was called the Parent Social Rating Scale in the ECLS-K.

Learning) of the parent interview, and parents used the same frequency scale as teachers to report how often their child demonstrated the behaviors described. The Approaches to Learning Scale score is the mean rating on the six items included in the scale. A score was computed when the respondent provided a rating on at least four of the six items that composed the scale. Higher scale scores indicate that the child exhibited positive learning behaviors more often. The variable names, descriptions, value ranges, weighted means, and standard deviations for the fall kindergarten and spring kindergarten parent Approaches to Learning Scale scores are shown in table 7-25. The Approaches to Learning Scale has an internal consistency reliability estimate of .70 for the fall data collection and .72 for the spring data collection, as measured by Cronbach's alpha. Additionally, the item-level data for the parent-reported Approaches to Learning items are included in the data file along with the other parent interview data. Variable names for the item-level data from the fall parent interview begin with "P1," and variable names for the spring parent interview begin with "P2."

Table 7-25. Parent-reported Approaches to Learning Scale variable names, descriptions, value ranges, weighted means, and standard deviations: School year 2010–11

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1PRNAPP	X1 Parent Report Approaches to Learning	13,220	1–4	3.18	0.477
X2PRNAPP	X2 Parent Report Approaches to Learning	13,240	1–4	3.13	0.489

NOTE: Fall kindergarten estimates weighted by W1P0. Spring kindergarten estimates weighted by W2P0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

Mean scores for the parent-reported Approaches to Learning Scale are presented by data collection round and child characteristics in table 7-26.

Table 7-26. Mean parent-reported Approaches to Learning Scale scores, by data collection round and child characteristics: School year 2010–11

Characteristic	Fall 2010			Spring 2011		
	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	13,220	3.18	0.477	13,240	3.13	0.489
Sex						
Male	6,800	3.12	0.488	6,810	3.06	0.494
Female	6,430	3.24	0.456	6,440	3.20	0.473
Race/ethnicity						
White, non-Hispanic	6,830	3.23	0.442	6,700	3.18	0.455
Black, non-Hispanic	1,700	3.16	0.500	1,480	3.14	0.516
Hispanic	2,980	3.07	0.516	3,140	3.03	0.524
Asian, non-Hispanic	880	3.10	0.507	1,100	3.03	0.521
Hawaiian, Other Pacific Islander, non-Hispanic	60	3.25	0.471	70	3.10	0.472
American Indian/Alaska Native, non-Hispanic	110	3.30	0.411	100	3.15	0.400
More than one race, non-Hispanic	660	3.22	0.450	650	3.16	0.472
School type						
Public school	11,400	3.17	0.481	11,460	3.12	0.492
Private school	1,820	3.26	0.435	1,670	3.22	0.447

NOTE: Fall kindergarten estimates weighted by W1P0. Spring kindergarten estimates weighted by W2P0. The unweighted sample size is the number of cases with a valid score rounded to the nearest 10. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2010 and spring 2011.

REFERENCES

Adler, C. Ralph. (Ed.). (2003). *Put Reading First: The Research Building Blocks for Teaching Children to Read.* 2nd Ed. Retrieved June 20, 2012 from http://lincs.ed.gov/publications/pdf/PRFbooklet.pdf.

Cohen, J. (1988). *Statistical Power for the Behavioral Sciences*, 2nd Ed. Hillsdale, NJ: Erlbaum.

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.

Diamond, A. Executive Functions. *Annual Review of Psychology, 64*: 135–168,

Dolch, E. W. (1948). *Problems in Reading*. Champain, IL: The Garrard Press.

Duncan, S. E., and De Avila, E. A. (1998). *pre*LAS *2000 Cue Picture Book English Form C*. Monterey, CA: CTB/McGraw-Hill Companies, Inc..

EducationWorld®. (n.d.) National science standards, grades K-4. *EducationWorld* (online education resource). Retrieved from http://www.educationworld.com/standards/national/science/k_4.shtml.

Dorans, N. J., and Kulick, E. (2006). Different item functioning on the Mini-Mental State Examination: An application of the Mantel-Haenzel and standardization procedures. *Medical Care, 44*(11) Suppl3: S107–S114.

Holland, P.W., and Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. (ETS Research Report No. 86-31). Princeton, NJ: ETs.

International Reading Association and the National Association for the Education of Young Children. (2008). *Learning to Read and Write: Developmentally Appropriate Practices for Young Children, part 4: Continuum of Children's Development in Early Reading and Writing.* A joint position of the International Reading Association (IRA) and the National Association for the Education of Young Children (NAEYC). Young Children, *53*(4):30–46. Retrieved June 20, 2012 from http://www.naeyc.org/files/naeyc/file/positions/PSREAD98.PDF .

International Reading Association. (1998). *Phonemic Awareness and the Teaching of Reading: A Position Statement from the Board of Directors of the International Reading Association.* Newark, Delaware: Author. Retrieved May 14, 2013 from http://www.reading.org/downloads/positions/ps1025_phonemic.pdf.

Kirsch, I. S., Jungblut, A., Jenkins, L., and Kolstad, A. (1993). *Adult Literacy in American: A First Look at the Results of the National Adult Literacy Survey* (NCES 1993-275). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Lee, J., Grigg, W. S., and Dion, G. S. (2007). *The Nation's Report Card: Mathematics 2007* (NCES 2007–494). Statistical Analysis Report. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

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.

Mather, N., and Woodcock, R.W. (2001). *Examiner's Manual. Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside Publishing.

Mislevy, R. J., and Bock, R. D. (1982). Bilog: Item analysis and test scoring with binary logistic models. [Computer program]. Mooresville, IN: Scientific sofrware.

Mislevy, R. J., Johnson, E. G., and Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*: 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. Mooresville, IN: Scientific Software.

Muraki, E. J., and Bock, R. D. (1991). PARSCALE: *Parameter scaling of rating data* [computer program]. Chicago: Scientific Software, Inc.

National Assessment Governing Board. (2004a). *Mathematics Framework for the 2005 National Assessment of Educational Progress*. Washington, DC: U.S. Government Printing Office. Retrieved from http://www.nagb.org/publications/frameworks.htm.

National Assessment Governing Board. (2004b). *Science Framework for the 2005 National Assessment of Educational Progress. Washington, DC; U.S. Government Printing Office*. Retrieved from http://www.nagb.org/publications/frameworks.htm.

National Assessment Governing Board. (2008). *Reading Framework for the 2009 National Assessment of Educational Progress*. Washington, DC: U.S. Government Printing Office. Retrieved June 20, 2012 from http://www.nagb.org/publications/frameworks/reading09.pdf

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics.* (Book and E-Standards CD). Reston, VA: Author. Online version available from http://www.nctm.org/standards/content.aspx?id=16909

National Mathematics Advisory Panel. (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education. Retrieved May 14, 2013 from http://www.ed.gov/mathpanel.

National Research Council. (1996). *National Science Education Standards. Washington, DC: National Academies Press*. Retrieved from http://www.nsta.org/publications/nses.aspx.

Odlin, T. (1989). *Language Transfer: Cross-Linguistic Influence in Language Learning*, Cambridge (United Kingdom): Cambridge University Press.

Pianta, R.C. and Steinberg, M. (Eds.) (1992). *Teacher-Child Relationships and the Process of Adjusting to School.* San Francisco, CA: Jossey-Bass.

Pollack, J. M., Rock, D.A., Weiss, M.J., and Atkins-Burnett, S. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–62). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/ecls/.

Putnam, S. P., and Rothbart, M. K. (2006). Development of Short and Very Short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment, 87*(1): 103–113.

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

Snow, C. E. (2002). *Reading for Understanding: Toward a Research and Development Program in Reading Comprehension*. Santa Monica, CA: RAND. Retrieved from http://www.rand.org/pubs/monograph_reports/2005/MR1465.pdf.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Hagedorn, M. C., and Daly, P. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011),User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic Codebook, Public Version* (NCES 2015-074). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

U.S. Department of Education. (2007). *Report of the Academic Competitiveness Council*. Washington, DC: Author. Retrieved from http://www.ed.gov/about/inits/ed/competitiveness/acc-mathscience/report.pdf.

U.S. Department of Education. (n.d). *The Facts About…Science Achievement.* (Archived *No Child Left Behind* information sheet.) Retrieved from http://www.ed.gov/nclb/methods/science/science.html.

Vukelich, C., and Christie, J. F. (2004). *Building a Foundation for Preschool Literacy: Effective Instruction for Children's Reading and Writing Development.* Newark, DE: International Reading Association.

Yamamoto, K., and Mazzeo, J. (1992). Item Response Theory: Scale Linking in NAEP. *Journal of Education Statistics, 17*: 155–173.

Yen, W. M. (1984). Effects of local item dependence on the fit and equation performance of the three-parameter logistic model. *Applied Psychological Measurement, 2:* 125–145.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A Method of Assessing Executive Function in Children. *Nature Protocols, 1*: 297–301.

APPENDIX A:
ECLS-K:2011 FALL 2009 FIELD TEST REPORT

Early Childhood Longitudinal Study Kindergarten Class of 2010–11 (ECLS-K:2011)

Fall 2009 Field Test Report

December 2011

Karen Tourangeau
Christine Nord
Alberto Sorongon
Westat

Michelle Najarian
ETS

Gail Mulligan, Project Officer
National Center for Educational Statistics

NOTE: The appendixes to the ECLS-K:2011 Fall 2009 Field Test Report are not included in this appendix.

TABLE OF CONTENTS

TABLE OF CONTENTS—CONTINUED

TABLE OF CONTENTS—CONTINUED

TABLE OF CONTENTS—CONTINUED

LIST OF TABLES

LIST OF TABLES—CONTINUED

LIST OF TABLES—CONTINUED

LIST OF FIGURES

LIST OF EXHIBITS

This page intentionally left blank.

1. INTRODUCTION AND BACKGROUND

This chapter provides basic information about the fall 2009 field tests conducted for the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011). It begins with an overview of the ECLS-K:2011 and of the design of the survey instruments and assessments, followed by a description of the field tests' purpose and field test materials. Later chapters provide information about the sample design (chapter 2), data collection and training (chapter 3), assessor feedback and recommendations (chapter 4), and the analysis of the field test results (chapters 5 and 6).

1.1 Study Overview

The ECLS-K:2011 is new study sponsored by the National Center for Education Statistics (NCES) within the Institute of Education Sciences, U.S. Department of Education and conducted by Westat. During the 2010–11 school year, approximately 20,700 kindergartners in 900 public and private schools across the nation will be selected to participate in the study. The ECLS-K:2011 will gather information from multiple sources including parents, teachers, schools, and care providers. Trained assessors will administer direct assessments of children's reading, mathematics, and science skills and of the children's executive functioning and working memory, which are associated with early learning.

The ECLS-K:2011 is the third in a series of longitudinal studies sponsored by the U.S. Department of Education that examine child development, school readiness, and early school experiences. It shares many of the same goals as its predecessors, the ECLS-K and the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), but also advances research possibilities by providing updated information and addressing recent changes in education policy.

- Like its predecessors, the ECLS-K:2011 will provide a rich and comprehensive source of information on children's early learning and development, transitions into kindergarten and beyond, and progress through school for a new cohort of children.
- The ECLS-K:2011 will provide data relevant to emerging policy-related domains not measured fully in previous studies.
- Coming more than a decade after the inception of the ECLS-K, the ECLS-K:2011 will allow cross-cohort comparisons of two nationally representative kindergarten classes experiencing different policy, educational, and demographic environments.

The longitudinal nature of the study will enable researchers to study cognitive and physical growth and socioemotional status, as well as relate trajectories of growth and change to variation in home and school experiences in the early grades. Ultimately, the ECLS-K:2011 data set will be used by policymakers, educators, and researchers to consider the ways in which children are educated in our nation's schools and to develop more effective approaches to education.

1.2 Design of ECLS-K:2011 Survey Instruments and Assessments

During the ECLS-K:2011 design phase, staff thoroughly reviewed the existing ECLS-K and ECLS-B nonassessment survey instruments (the school administrator, teacher, and parent interviews) to identify items that were appropriate for a new cohort of kindergartners and a changed education environment. Items that were still appropriate were brought forward to be included in the new ECLS-K:2011. These items will allow comparisons to be made between two cohorts of kindergarten classes that entered school more than a decade apart. Items that were deemed no longer relevant were dropped or revised. Through discussions with NCES, experts on the Technical Review Panel (TRP), and reviews of recent research, new areas of interest to NCES and to the research community were added. These new areas of interest will help the study provide data relevant to emerging areas of interest. The nonassessment survey instruments did not require field testing because the new items were drawn from existing studies or were very similar to items previously used.

The ECLS-B and ECLS-K cognitive assessment batteries were also thoroughly reviewed during the design phase using current state and national standards as guides. New items were developed when necessary, such as for executive functioning, to capture important content areas in the frameworks. For the ECLS-K:2011, NCES funded the development of a measure of Spanish-speaking children's reading skills and knowledge in their native language, the Spanish Early Reading Skills (SERS) assessment, and in English, the English Basic Reading Skills (EBRS) assessment. All of the items and assessments proposed for the ECLS-K:2011 were reviewed by measurement specialists and content area specialists (i.e., science experts reviewed all of the science items in each of the grades, reading experts reviewed the reading items, mathematics experts reviewed the mathematics questions, and experts in early reading skills and bilingual assessments reviewed the EBRS and SERS items). Based on the comments of the expert reviewers, the assessment items were revised and field test assessment materials developed.

In addition to the direct child assessments, the ECLS-K included Academic Rating Scales (ARS) completed by sampled children's teachers in language and literacy, mathematics, and science

(beginning in grade 3). The ARS scales were designed both to overlap and to augment the information gathered through the direct cognitive assessments. Most importantly, the ARS included items designed to measure both the process and products of children's learning in school, whereas the direct cognitive battery was more limited. Because of time and space limitations, the direct cognitive battery was less able to measure the process of children's thinking, including the strategies children use to read, solve mathematical problems, or investigate a scientific phenomenon. The ARS scales were successfully collected in the ECLS-K and will be repeated in the ECLS-K:2011.

The ECLS-K:2011 cognitive assessment battery consists of the following:

- Direct assessments of language and literacy, mathematics, and science, including a science assessment designed for younger children;
- Direct and indirect measures of executive function;
- Teacher ARS in language and literacy, mathematics, and science;
- SERS and EBRS assessments.

As they are established measures, the executive function assessments and the language and literacy and mathematics ARS scales required no field testing. The remaining ECLS-K:2011 cognitive assessments required field testing to obtain the psychometric properties of the item pools.

The ECLS-K:2011 may also include screening sampled children's hearing and vision. The National Center for Communication and Other Hearing Disorders and the National Eye Institute provided funds for a feasibility study during the field test of screening children's hearing and vision for possible inclusion in the national study. In this report, we describe the fielding of this feasibility study but do not provide a recommendation about whether auditory screening should be included in the national data collection. The data collected during the feasibility study were sent to the two funding organizations so that they could review the quality and decide whether to fund a national data collection.

1.3 Purpose of Field Tests

In fall 2009, two field tests were conducted to test the direct assessments, science ARS, and data collection procedures prior to the national study. These field tests are referred to as the English field test and the Spanish field test. These field tests served as the primary vehicle for (1) estimating the

psychometric parameters of all items in the assessment battery item pool, (2) producing psychometrically sound and valid direct and indirect cognitive assessment instruments, (3) assessing the feasibility of screening children's vision and hearing for the national collection, and (4) obtaining valid assessments for both an *English* reading score for Spanish-speaking children and an assessment of their early reading skills (e.g., letter recognition and sounds) *in Spanish.*

1.3.1 Purpose of the Fall 2009 English Field Test (EFT)

There were three goals for the fall 2009 English field test (EFT). The first goal was to collect cognitive assessment data for the development of the kindergarten assessments to be used in fall 2010 and spring 2011, as well as those to be used in the first- and second-grade data collections, to assess the cognitive development of children in the national sample. The field test provided the item statistics that will guide development of the battery of instruments to assess children's cognitive development in the domains of language and literacy, mathematics, and science.

The second goal of the EFT was to examine the feasibility of screening children's hearing and vision in a school setting. Health technicians screened hearing and vision on participating children in the EFT.

The third and final goal was to collect child rating data from teachers for the development of an ARS of children's science skills. Two teachers at each grade were asked to complete the Science ARS for five children in their classrooms.

As described in chapter 2, the EFT was conducted with a large purposive sample of approximately 3,000 kindergarten, first-grade, second-grade, and third-grade children. A purposive sample of 50 schools across five geographic areas was selected in order to get a range of urban, suburban, and rural schools, as well as public and private schools.

In each school, the goal was to assess an average of 63 children: 18 kindergartners, 18 first-graders, 18 second-graders, and 9 third-graders. In each school more children were sampled to account for children who were absent or withdrawn on the assessment days or for whom parents would not give permission to participate. Most EFT schools required signed parent permission forms in order to assess children, making it more difficult to conduct assessments with the desired number of children. In addition, there were some sampled schools with too few eligible children in each grade and other schools that had

more than enough. In an attempt to reach the overall goal of assessing 3,150 children, larger schools were asked to allow us to assess more than 63 children.

1.3.2 Purpose of the Fall 2009 Spanish Field Test (SFT)

The fall 2009 Spanish field test (SFT) served as the primary vehicle for (1) estimating the psychometric parameters of all items in the EBRS and SERS, (2) evaluating the content validity of the EBRS and SERS items, and (3) producing psychometrically sound and valid direct assessment instruments of English basic reading skills and Spanish early reading skills.

As described in chapter 2, the SFT was conducted with just over 1,100 Spanish-speaking kindergartners. A purposive sample of 50 schools in five geographic areas was selected based on the estimated percentage of the population that spoke Spanish.

In each school, the goal was to assess 24 Spanish-speaking kindergartners. As with the EFT, some of the sampled schools were smaller with too few children in this target population. Therefore, larger schools were asked to allow larger numbers of their children to participate. Since the burden of the SFT was not as great as the EFT, the impact on participation rates of schools requiring signed parent permission forms was not as great as it was with the EFT.

1.4 Field Test Data Collection

The field tests were conducted concurrently in fall 2009 to replicate the approximate time of the school year when the ECLS-K:2011 kindergarten data collection is scheduled to take place (fall 2010). Table 1-1 shows the overall fall 2009 field test activities and the schedule on which they were conducted.

Table 1-1. Fall 2009 field test activities and schedule

Activity	Time period
Prepare assessment materials	February–July 2009
Recruit assessors	May–July 2009
Develop training program	February–August 2009
Conduct training sessions	August 2009
Recruit schools	August–November 2009
Data collection	August–November 2009
Data preparation	October–December 2009
Data analysis	December 2009–January 2010

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

1.4.1 EFT Data Collection

The EFT had three components: 1) a direct child assessment of children in kindergarten through third grade; 2) an indirect assessment, the Science ARS, completed by teachers; and 3) a feasibility study of screening children's hearing and vision in a school setting.

EFT Direct Child Assessments

For each cognitive domain, the EFT assessments included items with a range of difficulty that helped to identify those children who are just beginning to develop skills in that domain, those who have solid skills, and those who are highly proficient or advanced. The items for each cognitive domain included some items that only about 5 percent of the children would be expected to be able to answer as well as some items that about 95 percent of the children would be expected to be able to answer. Although the EFT required that some children were administered items that may be well beyond their capability, the result of the field testing is a national assessment that is tailored to sampled children's skills. The EFT direct child assessments were administered in one-on-one sessions by trained assessors using easels for the stimulus items and their instructions and recording children's responses on score sheets.

Science ARS

As was done during field testing on ECLS-K study, six teachers at each school (two each in kindergarten, first, and second grade) were asked to complete the short ARS about children's skills in science. Teachers were given a package of five Science ARS forms and a cover letter. Each teacher was instructed to complete an ARS on five children in her classroom: the highest achieving child, the lowest achieving child, and three children with average achievement. Team leaders collected completed ARS forms while they were at the school on the days of the assessments or left shipping materials for the school coordinator to return them to Westat.

Hearing and Vision Screening

The hearing and vision screenings were conducted by health technicians using a battery of equipment at a hearing and vision screening station set up in each school. The ideal location for the screening station was separate from the assessment area, free of noise and visual distractions, and at least 9 feet long to accommodate one vision test. Two different vision tests were conducted during the field test, one measuring visual acuity, and the other refractive error and the shape of the cornea. Three different hearing tests were conducted: otoscopy, which examines the children's ear canal and eardrums; tympanometry, which tests the mobility of the eardrum; and audiometry, which measures acoustic reflex and is used to detect hearing loss.

1.4.2 SFT Data Collection

The SFT was also a direct assessment conducted one-on-one with children using easels for the stimulus items and recording children's responses on score sheets. The SFT direct assessment was conducted with a warm-up easel and an English Basic Reading Skills/Spanish Early Reading Skills (EBRS/SERS) assessment easel. The warm-up easel contained two simple practice items that were designed to give the child an idea of the types of questions to be asked and help him/her transition into the assessment. The EBRS/SERS assessment easel was further divided into two sections: the EBRS section was administered directly after the warm-up easel, followed by the SERS section.

This page intentionally left blank.

2. FIELD TEST SAMPLE DESIGN

2.1 Field Test Samples

One hundred schools were initially asked to participate, 50 for each field test. To identify the appropriate number of schools to participate, 20 public school districts and Catholic dioceses were contacted by trained, experienced school recruiters. As districts/dioceses agreed to participate in the study, school recruiters contacted schools throughout the field period to recruit enough schools to achieve the required number of completed assessments.

2.2 EFT Sample

The EFT design called for a purposive sample of 50 schools to obtain 3,150 completed child assessments and 300 completed Science ARS'. The EFT was conducted in five states: Colorado, Maryland, Ohio, Pennsylvania, and South Carolina.

Although the sample of schools was purposive, an effort was made to select and recruit both public and private schools with a variety of characteristics, such as percent of minority children, locale (i.e., rural, urban, suburban), and enrollment size. To ensure that the sample included both high achieving and low achieving schools, school median income rankings were used as a proxy for school achievement rankings since the two variables are highly correlated. Using data from the 2000 Census long form, the median household income was calculated at the five-digit ZIP code level. It was then attached to each school based on the ZIP code of the location address where possible, and on the mailing address otherwise.

During school recruitment, 29 school districts and 94 schools were contacted. A total of 37 public, Catholic, and other private schools agreed to participate in the fall 2009 EFT. Schools that refused to participate generally cited the start of the school year and the study burden as the two main reasons for refusal. Participating schools were paid a $5 honorarium for each completed child assessment. Table 2-1 presents the characteristics of the participating EFT schools.

Table 2-1. Characteristics of schools participating in the fall 2009 English field test

Characteristic	Number	Percent
Total schools	37	100
School sector		
Public	32	86
Private	5	14
Catholic	4	11
Other religious	1	3
Nonsectarian	0	0
Percent minority		
1–9	9	24
10–29	18	49
30–49	4	11
50 percent or more	6	15
Unknown	0	0
Community type		
Rural or small town	5	14
Urban fringe or large town	25	68
Central city	7	19
Total school enrollment		
1–79	0	0
80–200	1	3
201–349	8	22
350–500	10	27
501 or more	18	49

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

Additionally, more than two-thirds of the EFT schools required signed parent permission forms before assessments could be conducted (table 2-2).

Table 2-2. Type of consent required for the fall 2009 English field test, by school type

	Signed parent permission form		Notification only	
School type	Number	Percent	Number	Percent
Total schools	25	68	12	32
Public	23	92	10	83
Catholic	2	8	1	8
Other religious	0	0	1	8
Nonsectarian	0	0	0	0

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

Table 2-3 presents the characteristics of the 2,997 children who completed assessments.

Table 2-3. Characteristics of children participating in the fall 2009 English field test

Characteristic	Number	Percent
Total children	2,997	100
Gender		
Male	1,517	51
Female	1,470	49
Unknown	10	0
Race[1]		
American Indian or Alaska Native	16	1
Asian, Native Hawaiian, or Pacific Islander	75	3
Black non-Hispanic	340	11
White, non-Hispanic	2,225	74
Hispanic	299	10
Other	29	1
Unknown	13	<1
Age on 11/30/09		
4–5 years	682	23
6–7 years	1,682	56
8–9 years	617	21
10 years and older	1	0
Unknown	15	1
Ability level[1]		
Above grade	736	25
On grade	1,607	54
Below grade	570	19
Unknown	84	3

[1] As determined by the school.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

Additionally, 2,101 children participated in the vision screening; of those, 416 children completed the hearing screening.

2.3 SFT Sample

Approximately 50 schools in five geographic areas were selected for the SFT based on the percentage of the population that spoke Spanish in those areas. The goal was to assess approximately 1,200 Spanish-speaking kindergartners. The field test was conducted in four states: California, Florida, New Mexico, and Texas. As described above, school recruiters first contacted school districts and then schools; 36 schools agreed to participate. The SFT schools determined how to select participating

children following guidance from the school recruitment staff. Only one charter school required signed parent permission forms in order for children to participate. The remaining 35 schools required only parental notification.

Table 2-4 presents the characteristics of the 36 schools participating in the fall 2009 Spanish field test.

Table 2-4. Characteristics of schools participating in the fall 2009 Spanish field test

Characteristic	Number	Percent
Total schools	36	100
School sector		
Public	32	86
Private	4	14
Catholic	3	11
Other religious	0	3
Nonsectarian	1	0
Percent minority		
1–9	0	0
10–29	0	0
30–49	0	0
50 percent or more	35	97
Unknown	1	3
Community type		
Rural or small town	5	14
Urban fringe or large town	12	33
Central city	19	53
Total school enrollment		
1–79	0	0
80–200	1	3
201–349	1	3
350–500	8	22
501 or more	25	69
Unknown	1	3

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Spanish field test, fall 2009.

Table 2-5 presents the characteristics of the 1,123 children who completed assessments.

Table 2-5. Characteristics of children participating in the fall 2009 Spanish field test

Characteristic	Number	Percent
Total children	1,123	100
Gender		
Male	559	50
Female	563	50
Unknown	1	0
Race[1]		
American Indian or Alaska Native	2	0
Asian, Native Hawaiian, or Pacific Islander	6	1
Black non-Hispanic	3	0
White, non-Hispanic	7	1
Hispanic	1,102	98
Other	3	0
Unknown	1	0
Age on 11/30/09		
4–5 years	887	79
6 years	231	21
7 years	2	0
8 years	1	0
Unknown	2	0
Ability level[1]		
Above grade	137	12
On grade	523	47
Below grade	331	29
Unknown	132	12

[1]As determined by the school.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Spanish field test, fall 2009.

This page intentionally left blank.

3. DATA COLLECTION

This chapter describes data collection, organization of the field work, school recruitment, the two major components of data collection—direct assessments and teacher ratings, field staff communication, and quality assurance.

3.1 Field Organization

The fall 2009 field test data collection period began on August 24, 2009, and ended on November 20, 2009. To maximize the data collection efficiency, the fall 2009 field test sample was organized into 10 work areas. Each work area contained 10 field test schools.

The English field test teams, consisting of one team leader, three assessors, and two health technicians, were each assigned to one work area and responsible for assessing children and conducting auditory screening at the schools in that work area. Table 3-1 identifies the work areas.

Table 3-1. Fall 2009 English field test work areas

Work area number	Work areas
E 1	Denver, CO
E 2	Anne Arundel, MD
E 3	Lake, OH
E 4	Montgomery, PA
E 5	Lexington, SC

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

The Spanish field test teams, consisting of one team leader and one assessor were each assigned to one work area and responsible for assessing children at the schools in that work area. Table 3-2 identifies the work areas.

Table 3-2. Fall 2009 Spanish field test work areas

Work area number	Work areas
S 1	Fresno, CA
S 2	Broward, FL
S 3	Bernalillo, NM
S 4	Bexar, TX
S 5	El Paso, TX

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Spanish field test, fall 2009.

3.2 Field Staff Training

Four training programs were developed and conducted for the fall 2009 field tests: school recruiter training, team leader training, assessor training, and health technician training. All training programs included a home study training package followed by an in-person training program. The focus of all training programs was to provide trainees with sufficient practice to be proficient with their tasks.

The importance of keeping all information about the participating schools and respondents confidential was stressed in each of the training manuals and during all in-person training sessions. Furthermore, all Westat field employees are required to view the corporate Data Security module and sign an Affidavit of Nondisclosure as a condition of employment.

3.2.1 School Recruiter Training

Seven experienced school recruiters were trained to recruit field test schools. The school recruiter in-person training session was held August 12–14, 2009, in Rockville, Maryland. The home study included the School Recruiter Manual, exercises, and copies of the advance packages sent to states and districts. In the home study, recruiters participated in extensive practice addressing respondent concerns and objections, including tape-recorded practice and telephone role plays with fellow recruiters, and reviewed all the materials used to contact and schedule schools. The in-person training provided additional practice on recruiting schools into the study, scheduling assessments for each work area, notifying parents and obtaining their consent, and negotiating for space to conduct the direct assessments the hearing and vision screenings (EFT only). (See appendix A for the School Recruiter In-Person Training Agenda.)

3.2.2 Team Leader Training

English Field Test

Five team leaders were trained to supervise the teams of assessors conducting the EFT direct assessments. A Field Test Team Leader Manual was sent to all team leaders prior to attending the in-person training session. The manual provided an overview of the EFT and described their responsibilities

as team leaders. Team leaders also received the sections of the school recruiter home study training focusing on addressing respondent questions and concerns.

The team leader in-person training was held on August 17, 2009. The team leader training (see appendix B for the EFT Team Leader Training Agenda) consisted of lectures and exercises covering their roles and responsibilities including preassessment contacts with schools, managing and conducting assessments, and postassessment activities of packing and mailing completed materials to the home office. Other topics included motivating the team and reporting status on activities. (See appendix B for the EFT Team Leader Training Agenda.)

Spanish Field Test

Five team leaders were trained to supervise the teams of assessors conducting the SFT. The training program was similar to the EFT team leader training described above but modified to reflect the differences in the SFT assessments. The SFT team leader in-person training was held on September 9, 2001. (See appendix C for the Spanish Field Test Team Leader Training Agenda.)

3.2.3 Assessor Training

English Field Test

The assessor in-person training held August 18–21, 2009, in Rockville, Maryland. The home study package, including the Field Test Assessor Manual, was sent to all assessors prior to attending the in-person training session. Fifteen assessors and five team leaders attended the training. (The EFT Assessor Training Agenda is included in appendix D.)

The majority of the assessor training was devoted to the proper techniques for administering the EFT direct assessments, including using the associated easels and score sheets, gesturing norms, and standardized approaches for administering the assessment battery. Additional topics included general procedures for working with young children, using data collection materials, setting up the assessment space, and getting each child to and from the classroom to the assessment area.

The basic training approach for each cognitive assessment was a group lecture and demonstration of one of the assessment forms, followed by individual practice and role-play dyads on the

second form of each of the assessment domains. After instruction on all of the cognitive domains was completed, the assessors were paired with each other. They took turns administering the assessment battery to each other while being observed by one of the training staff. Observers provided feedback to the assessors on administering and scoring the assessment items and strategies for improving rapport with children.

Spanish Field Test

SFT assessor training followed the EFT program outlined above with the in-person training session held September 9–11, 2009, in Rockville, Maryland. Five assessors and five team leaders attended the training. (The SFT Assessor Training Agenda is included in appendix E.)

3.2.4 Health Technician Training

The health technician in-person training was held August 17–21, 2009, in Rockville, Maryland. Ten health technicians attended the training. Prior to attending an in-person training session, health technicians received a home study packet that included their procedural manual and an exercise regarding the information included in the manual.

The health technician training included descriptions and demonstrations of each piece of equipment used for auditory screening, practice using each piece of equipment, setting up and calibrating equipment, packing up and caring for equipment, and how to work with young children. (See appendix F for the Health Technician In-Person Training Agenda.) The basic training approach was a trainer demonstration of the setup and use of each piece of equipment to all technicians followed by trainee practice with the equipment. A significant portion of training time was allotted for pairs of technicians to practice examining each other using the equipment while the trainers observed, gave feedback, and assisted each pair of technicians. Vision equipment and exams were covered during the first 2 days of training, and hearing equipment and exams were discussed during the last 2 training days. Children were brought in on the last training day to allow the technicians to practice examining child test subjects prior to beginning data collection in the schools.

A small subset (4 of the 10 health technicians) received instruction on the use of an additional piece of equipment still in the experimental stage (the AMTAS) on the day prior to the start of training.

3.3 School Recruitment

School recruitment began in mid-August and continued through mid-November 2009. Schools were sent an advance packet of materials explaining the ECLS-K:2011 and providing information about the importance of the study (see appendix G for the School Advance Package materials). Using the School Contact Log (see appendix H for the School Contact Log), school recruiters then called each school to discuss participation. During the telephone call, recruiters:

- Explained the purpose of the fall 2009 field test;
- Identified a school coordinator to be a liaison between the school and the study;
- Negotiated parent notification and how parent consent would be obtained;
- Identified where the assessments would take place;
- Scheduled the assessment days; and
- Identified how to distribute the Science ARS to the participating teachers (EFT only).

The person designated as the school coordinator was usually a staff member (e.g., assistant principal, curriculum coordinator, or guidance counselor) designated by the school principal to serve as a liaison with the study. Sometimes the principal or a teacher fulfilled this role. School recruiters worked with school coordinators to identify an appropriate space for conducting the assessment, determine how children would be identified for the assessment, estimate the number of children by grade that could be assessed, manage distribution and collection of signed parental permission forms, and provide logistical information about the school location and parking. In the EFT, teachers were identified to complete the Science ARS and the location of the vision and hearing screening station was negotiated.

School recruiters reported this information to the home office to prepare and mail school coordinator, teacher, and parent packets to the schools. School coordinator packets contained a letter and a description of activities for which the coordinator was asked to assist along with other information about the ECLS-K:2011 (see appendix I for the School Coordinator package). Teacher packets contained a letter and other information about the ECLS-K:2011 (see appendix J for the Teacher package). For the EFT, teacher packets also contained the Science ARS and instructions for how to complete and return

them to the school coordinator. Parent packets contained a letter and general information about the study, as well as a parent permission form (see appendix K for the Parent package).

After school recruiters obtained cooperation from the schools, they continued to follow up with schools to plan assessment activities and receive updates on the status of parent permission forms as necessary. If the school required signed parent permission forms, school recruiters used the Parent Permission Form Tally Sheet to record counts of returned parent permission forms. School coordinators were asked to use a Parent Permission Tracking Form (PPTF) (see appendix L for an example of the Parent Permission Tracking Form) to record information for each participating child. There was one color-coded form for each grade: kindergarten (yellow), first grade (peach), second grade (blue), and third grade (green). The PPTF included a column for "Assessment Window" where school coordinators indicated an ideal day and/or time of day for assessing each child. These follow-up contacts continued up to a week prior to the first scheduled assessment date.

3.4 Preassessment Call to the School Coordinator

Approximately 1 week prior to the scheduled assessment date at a school, the team leader contacted the school coordinator to confirm assessment logistics including assessment dates and locations, status of parent permission forms, numbers of participating children, and school check-in procedures.

3.5 Meeting With the School Coordinator

On the first day of assessments at a school, team leaders met with school coordinators to finalize assessment plans for the day and verify information received in the preassessment call. School coordinators identified teachers and classrooms from which participating children would be drawn. If parent permission forms had been received for more children than could be assessed during the scheduled assessment period, team leaders also reviewed specific procedures to identify which children would be assessed.

3.6 Assessment Logistics

The space available for assessments in the schools varied considerably. In some schools, libraries or multipurpose rooms were available to conduct assessments; in other schools, assessments

were conducted in offices, hallways, or other facilities physically isolated from the main school area. Additionally, some schools had different rooms available at different times of the day or week, and the team had to change locations each day or during the day as the available space changed. Establishing the assessment space and moving children back and forth between the classroom and the assessment area presented a major logistics challenge in most schools. Differences in assessment locations and child retrieval procedures directly affected the time available for conducting the assessments. When assessment locations, such as portable classrooms, were isolated from the classrooms, retrieving the children could be time consuming, particularly in inclement weather. Finding appropriate space for the vision and hearing screenings was problematic throughout the field period. The vision screenings required low lighting and a distance of at least 9 feet to accommodate the visual acuity testing, and the hearing exams required low noise levels. In addition, there were two technicians in each school, so two suitable spaces had to be identified (either in one large room or two smaller rooms). Technicians could not close the door of the screening room if doing so meant the technician and child would be alone in the room, so noise from adjoining classrooms, lunchroom, and hallways frequently rendered the environment less than ideal for the hearing screenings. Fluorescent lighting and small rooms also created problems for the effective administration of the vision exams.

A Teacher Sign Out Sheet was posted just inside the classroom door and used to keep track of the children being assessed from a particular teacher's class. Assessors used these sheets to record each child's name, assessment room, and the time the child left the classroom and returned, and then added their initials.

3.7 Student and Teacher Identification

Child-level materials were labeled with child ID numbers. Child ID numbers included the school ID number and the unique two-digit child number (e.g., F00101). The Science ARS' were labeled with teacher ID numbers. Teacher ID numbers included the school ID, the letter "T," and two digits from 01 through 012 (e.g., E101T01, E101T02, etc.).

3.8 Conducting the Assessment

The EFT assessment materials consisted of a set of assessment easels in the three cognitive domains with each assessment easel containing a reading assessment and either a mathematics or science assessment. The field test items were grouped by grade with a set of easels for children in Kindergarten-first grade and another set of easels for children in second and third grade. For each grade group, there

were four field test versions of the reading assessment and two versions each of the mathematics and science assessments. Because the total number of items to be field tested was too large to be administered to any single child, a block and spiral design was used. Different versions for each cognitive area were created in order to prevent children from becoming fatigued due to the number of items in each test. That is, the items in each domain were first split into subsets of items of roughly equivalent difficulty; next, the subsets were arranged into easels that contained one subset for each of two domains, as depicted in table 3-3. A single easel was administered to each child.

For each grade group, K–1 or 2–3, eight different assessment easels were administered, resulting in a total of 16 assessment easels across the two grade groups with each easel labeled with a number and color. Each assessment easel had its own score sheet on which to record children's responses. The easels were created as flip books so that when opened, the assessor, who sat facing the child, viewed one side while the child viewed the other side. For each item, the child side of the easel presented the stimulus (e.g., an image and/or text) and on the assessor side of the easel presented the instructions on how to administer the item. The easels were color coded by grade and version to ensure the correct version of the assessment was administered to each child. The assessor recorded the child's responses on the appropriate score sheet. The score sheet covers were color coded by grade and version to match the corresponding easel. The score sheet included space to record the child ID, date, assessor ID, assessment status code, and any comments about the child's assessment. Each side of the score sheet contained the response grids associated with the cognitive domains for that easel (e.g., for the easel 1 score sheet, response grids for reading 1 were on pages 1 and 2 and the response grids for mathematics 1 were on pages 3 and 4).

Table 3-3. Components of the fall 2009 English field test assessment easels

Easel	Instruments	
Grade K–1 group		
Easel 1—Daffodil	Reading 1 (65 items)	Mathematics 1 (55 items)
Easel 2—Orange	Reading 2 (65 items)	Mathematics 2 (55 items)
Easel 3—Red	Reading 3 (65 items)	Science 1 (55 items)
Easel 4—Pale Yellow	Reading 4 (65 items)	Science 2 (55 items)
Easel 5—Goldenrod	Mathematics 1 (55 items)	Reading 1 (65 items)
Easel 6—Salmon	Mathematics 2 (55 items)	Reading 2 (65 items)
Easel 7—Yellow Crayons	Science 1 (55 items)	Reading 3 (65 items)
Easel 8—Red Crayons	Science 2 (55 items)	Reading 4 (65 items)
Grade 2–3 group		
Easel 9—Bright Blue	Reading 1 (70 items)	Mathematics 1 (55 items)
Easel 10—Lime	Reading 2 (70 items)	Mathematics 2 (55 items)
Easel 11—Bright Green	Reading 3 (70 items)	Science 1 (55 items)
Easel 12—Pale Blue	Reading 4 (70 items)	Science 2 (55 items)
Easel 13—Pale Green	Mathematics 1 (55 items)	Reading 1 (70 items)
Easel 14—Pale Purple	Mathematics 2 (55 items)	Reading 2 (70 items)
Easel 15—Blue Crayons	Science 1 (55 items)	Reading 3 (70 items)
Easel 16—Green Crayons	Science 2 (55 items)	Reading 4 (70 items)

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

The assessor conducted the assessment by reading the item script from the assessor side of the easel while the child looked at the stimulus on the other side and scoring the child's responses on the score sheet. The assessor's side also contained instructions on prompting children, proper gesturing, and scoring that were not read to the child.

The SFT assessment materials were also flip book easels with children's responses recorded on score sheets. The SFT score sheet was printed with the EBRS response grids on the front and the SERS response grids on the back. There was also space to record the child ID, date, assessor ID, assessment status code, and any comments about the child's assessment.

Team leaders identified a central place in the assessment room where the Child Administration Records (CAR) (see appendix M for an example of the Child Administration Record), Parent Permission Tracking Forms, and score sheets were available for all team members. The CAR was prefilled with child IDs and booklet versions as a way to identify which assessment easel and score sheet to use.

During preassessment contacts and on assessment day, team leaders worked with school coordinators to ensure that basic child information, including name, grade, gender, race/ethnicity, date of birth, and ability level, was collected on the Parent Permission Tracking Form. At the beginning of the

assessment, assessors transcribed that information onto the CAR. If any information was found to be missing, team leaders consulted with school coordinators to obtain the information.

Descriptions for the race/ethnicity codes used were as follows:

- **AI/AN** American Indian or Alaska Native, not of Hispanic origin
- **AS** Asian, not of Hispanic origin
- **PI** Native Hawaiian or Pacific Islander, not of Hispanic origin
- **B** Black or African American, not of Hispanic origin
- **W** White, not of Hispanic origin
- **H** Hispanic origin of any race
- **O** Other (includes multiracial non-Hispanic)

The ability level codes were defined as follows:

- **AGL** Above grade level
- **GL** Grade level
- **BGL** Below grade level

3.9 Hearing and Vision Screenings

After each assessment was completed, assessors took the child to the next available hearing/vision screening station. Two health technicians on each team conducted the vision and hearing screenings at schools participating in the EFT. The location of the testing stations varied by school depending on the availability of suitable space. The modal assessment and screening schedule can be found in exhibit 3-1.

Exhibit 3-1. Assessments and hearing/vision exam schedule

Daily assessment schedule (9:15 am–3:15 pm)							
	Child Assessments				Hearing/Vision Exams		
Session	Assessor 1	Assessor 2	Assessor 3	Assessor 4	Session	HT 1	HT2
9:15-10:15	C1	C2			10:15-10:45	C1	C2
9:45-10:45			C3	C4	10:45-11:15	C3	C4
10:30-11:30	C5	C6			11:30-12:00	C5	C6
11:00-12:00			C7	C8	12:00-12:30	C7	C8
12:00-1:00	C9	C10			1:00-1:30	C9	C10
12:30-1:30			C11	C12	1:30-2:00	C11	C12
1:15-2:15	C13	C14			2:15-2:45	C13	C14
1:45-2:45			C15	C16	2:45-3:15	C15	C16
Note: *30 minutes provided for lunch*							

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

Children whose ECLS-K identification number ended in an odd number received the hearing examination first; those with even numbered IDs received the vision screening first. The Hearing/Vision Testing Station Form (see appendix N Hearing/Vision Testing Station Form) contained questions to identify children who should not be tested (such as those with a severe eye infection) and also provided information useful to researchers analyzing the results of the screening efforts. Children were screened following the protocols and technicians recorded results from each screening into a laptop. At the conclusion of each exam, additional information—such as the overall quality of the exam and any adverse conditions during the exam—was recorded on the child's Testing Station Form.

The following equipment was used to screen vision:

- Electronic Visual Acuity (EVA)—an electronic device incorporating a specially programmed Palm Pilot that communicates with a laptop computer and is used to test distance visual acuity.
- Retinomax K-plus—an objective refractometer, designed to measure refractive power, corneal shape, and pupil diameter.

The following equipment was used to screen hearing:

- Welch-Allyn Model 25020 otoscope—a small, hand-held instrument with a light that is directed through a funnel-like tip to illuminate the ear canal for examination.

- Quest Technologies Model BA-202-27 bioacoustic simulator and octave band monitor—performs two functions: 1) as a bioacoustic simulator, it is a kind of "dummy" ear that is used to check the calibration of the audiometer on a daily basis; and 2) as an octave band monitor, it is used to continuously measure the background noise levels in the audiometric test room.

- Cardinal Health Model GSI-39 Auto Tymp combination audiometer and tympanometer—as a tympanometer, it is used to evaluate the functional health of the middle ear system; as an audiometer, it is used to obtain air conduction thresholds.

- AMTAS (Automated Method for Testing Auditory Sensitivity) audiometry system—performs audiometric screenings that allow the subject to respond to tones using a "game" on a touch screen monitor rather than by merely raising a hand. The ECLS-K field test used a version of the AMTAS specifically designed for administration to children (the KIDTAS).

The initial protocol for the hearing screenings involved the use of insert earphones, otoscopy, and tympanometry on all children. This decision was reconsidered as a result of difficulties encountered by the technicians in inserting the earphones, probe, or scope during training. Hearing screenings in the EFT schools were consequently delayed while a new protocol was created, which eventually resulted in far fewer hearing exams during the field test than vision exams (416 hearing exams compared to 2,101 vision exams). The new protocol called for the use of over-the-ear headphones (or "cans") for children in kindergarten or first grade, and the use of insert earphones on children in grades 2 and 3. The use of tympanometry and otoscopy was also restricted to the older children in the study (second- and third-graders).

The vision and hearing equipment was calibrated at the beginning of each day before exams were conducted, and the results of daily calibration were recorded on a Daily Calibration Form. If the equipment was moved to a different location after it had been calibrated, the equipment was recalibrated. Health technicians were also responsible for disinfecting and preparing the equipment after each child examination. All of the equipment either utilized disposable covers or was sanitized with alcohol wipes.

At the conclusion of each day, data from each exam station was transferred to a single, school-specific flash drive to be shipped back to Westat for processing. In addition, each health technician was responsible for completing one Hearing/Vision Testing Feasibility Form (see appendix O for the Hearing/Vision Testing Feasibility Form) daily to document that day's exam issues. Unlike the Testing Station Form, which recorded information about a given child's exam, the Feasibility Form recorded overall conditions in the testing area for the day, such as excessive noise, lack of adequate space, or difficulty with equipment.

3.10 Production Results for Child Assessments

3.10.1 English Field Test

A total of 2,978 children were assessed in the fall 2009 English field test. Table 3-4 presents the number of direct assessments completed, by various school characteristics, including school sector, percent minority, community type, size of enrollment, and grade.

Table 3-4. English field test: Number of direct assessments completed, by school characteristics: 2009

Characteristic	Assessments	
	Number completed	Percent completed
Total direct assessments	2,978	100
School type		
Public	2,554	86
Private		
Catholic	407	14
Other religious	17	<1
Nonsectarian	0	0
Percent minority		
1–9	668	22
10–29	1,364	46
30–49	370	12
50 percent or more	490	17
Unknown	86	3
Community type		
Rural or small town	171	6
Urban fringe or large town	1,592	53
Central city	635	21
Unknown	580	19
Total school enrollment		
1–79	0	0
80–200	17	1
201–349	665	22
350–500	807	27
501 or more	1,489	50
Grade		
Kindergarten	905	30
First	846	30
Second	818	28
Third	409	14

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

Table 3-5 illustrates the number of hearing screenings completed, by various school characteristics, including school sector, percent minority, community type, size of enrollment, and grade.

Table 3-5. Number of hearing screenings completed, by school characteristics: 2009

Characteristic	Screenings	
	Number completed	Percent completed
Total hearing screenings	416	100
School type		
Public	416	100
Private		
Catholic	0	0
Other religious	0	0
Nonsectarian	0	0
Percent minority		
1–9	114	27
10–29	235	57
30–49	0	0
50 percent or more	0	0
Unknown	67	16
Community type		
Rural or small town	0	0
Urban fringe or large town	416	100
Central city	0	0
Total school enrollment		
1–79	0	0
80–200	0	0
201–349	129	31
350–500	127	30
501 or more	160	39
Grade		
Kindergarten	104	25
First	138	33
Second	152	37
Third	22	5

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

Table 3-6 illustrates the number of vision screenings completed, by various school characteristics, including school sector, percent minority, community type, size of enrollment, and grade.

Table 3-6. Number of vision screenings completed, by school characteristics: 2009

Characteristic	Screenings	
	Number completed	Percent completed
Total vision screenings	2,101	100
School type		
Public	1,793	85
Private		
Catholic	290	14
Other religious	18	1
Nonsectarian	0	0
Percent minority		
1–9	469	22
10–29	851	41
30–49	218	10
50 percent or more	477	23
Unknown	86	4
Community type		
Rural or small town	146	7
Urban fringe or large town	1,400	67
Central city	555	26
Total school enrollment		
1–79	0	0
80–200	18	1
201–349	485	23
350–500	540	26
501 or more	1,058	50
Grade		
Kindergarten	649	31
First	642	31
Second	537	25
Third	273	13

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

3.10.2 Spanish Field Test

A total of 1,115 kindergartners were assessed in the SFT. Table 3-7 illustrates the number of direct assessments completed, by various school characteristics, including school sector, percent minority, community type, size of enrollment, and grade.

Table 3-7. Spanish field test: Number of direct assessments completed, by school characteristics: 2009

Characteristic	Assessments	
	Number completed	Percent completed
Total direct assessments	1,115	100
School type		
Public	1,081	97
Private		
Catholic	26	2
Other religious	0	0
Nonsectarian	8	1
Percent minority		
1–9	0	0
10–29	0	0
30–49	0	0
50 percent or more	1,115	100
Unknown	0	0
Community type		
Rural or small town	304	27
Urban fringe or large town	337	30
Central city	474	43
Total school enrollment		
1–79	0	0
80–200	8	1
201–349	8	1
350–500	220	19
501 or more	879	79

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Spanish field test, fall 2009.

3.11 Production Results for Science Academic Rating Scales

The field tested Science ARS questions focused on topic areas in science that are typically a part of the curriculum in kindergarten, first, and second grades, with examples of relevant skills for each topic area. In order to capture the range of abilities children may have at each grade level, the examples of skills for each topic area included some high-level skills. For each item, teachers were asked to rate the child on the following **five-point scale** which reflects the degree to which a child has acquired and demonstrates the targeted skills, knowledge, and behaviors (exhibit 3-2).

Exhibit 3-2. Response scale for the Science Academic Rating Scale

Response anchor	Definition	Value
Not yet	Child has not yet demonstrated skill, knowledge, or behavior.	1
Beginning	Child is just beginning to demonstrate skill, knowledge, or behavior but does so very inconsistently.	2
In progress	Child demonstrates skill, knowledge, or behavior with some regularity but varies in level of competence.	3
Intermediate	Child demonstrates skill, knowledge, or behavior with increasing regularity and average competence but is not completely proficient.	4
Proficient	Child demonstrates skill, knowledge, or behavior competently and consistently.	5
Not applicable or skill not yet taught	Skill, knowledge, or behavior has not been introduced in classroom setting.	–

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

If the skill, knowledge, or behavior had been introduced in the classroom, teachers were instructed to rate only the child's **current** achievement. If the skill, knowledge, or behavior had not been introduced in the classroom setting, teachers were instructed to select "not applicable or skill not yet taught."

In EFT schools, teachers of children in kindergarten, first grade, and second grade were each sent a packet of five Science ARS forms. Two teachers in each grade were asked to complete the Science ARS forms. The participating teachers were asked to choose five children from their class and rate the children's science skills and knowledge using the forms for this current grade level. Specifically, they were asked to rate their highest achieving student, their lowest achieving student, and three students with average achievement, regardless of whether these children were participating in the direct assessment. They were instructed to select children in their classrooms who matched the criteria and think about their skills and knowledge in science when completing the forms. Teachers indicated the child's ability level of the child on each form. As a result, Science ARS field test data were collected on children with different abilities.

School coordinators were mailed teachers packages to distribute to teachers before the scheduled assessments. The teacher packages included letter, five Science ARS, instructions for completing them, and an incentive check. Team leaders collected all completed ARS at the conclusion of the assessment period. Teachers who had not completed their ARS were asked to mail them to the home office. Team leaders followed up with telephone calls to the teachers to encourage them to complete the

forms. To ensure that enough completed rating forms were collected, schools that were more agreeable to the study were asked to determine if another teacher would each be willing to complete five rating forms. In the end, five schools agreed to allow additional teachers to complete the Science ARS. In total, these five schools yielded 18 additional teachers.

At least one set of five teacher rating forms was received from all of the participating EFT schools. A total 1,208 teacher questionnaires were collected: 423 kindergarten, 410 first grade, and 375 second grade. Table 3-8 presents the number of teacher questionnaires completed by various school characteristics, including type of school, percent minority, community type, school enrollment, and grade.

Table 3-8. Number of teacher rating forms completed, by school characteristics: 2009

Characteristic	Teacher rating forms	
	Number completed	Percent completed
Total teacher rating forms	1,208	100
School type		
Public	1,073	89
Private		
Catholic	120	10
Other religious	15	1
Nonsectarian	0	0
Percent minority		
1–9	275	23
10–29	585	48
30–49	138	11
50 percent or more	170	14
Unknown	40	3
Community type		
Rural or small town	60	5
Urban fringe or large town	710	59
Central city	203	17
Unknown	235	19
Total school enrollment		
1–79	0	0
80–200	15	1
201–349	213	18
350–500	305	25
501 or more	675	56
Grade		
Kindergarten	423	35
First	410	34
Second	375	31

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

3.12 Field Staff Communication

Weekly report calls between the field manager and team leaders were conducted throughout the field period to monitor status and identify issues. Typical report call topics included production, personnel issues, scheduling, assessment administration, and supplies. Team leaders were in daily communication with their team members and communicated results of the weekly report calls to them.

3.13 Quality Assurance

One of the critical aspects of quality assurance occurred during assessor training; only those assessors who were judged by the training team to have reached proficiency in administering the assessments were given field assignments.

Team leaders and assessors alike were trained to exchange Child Administration Records to verify data were transferred properly from the Permission Tracking Form. Additionally, team leaders were trained to scan score sheets and the Science ARS for completeness before packing them up for return to Westat.

To ensure that assessment protocols were properly administered, observers from the National Center for Education Statistics (NCES), the Education Statistics Services Institute (ESSI), and experienced Westat supervisors visited several schools to watch teams conduct assessments. Activities upon which the assessors were observed included:

- Correctly reading graphs and tables;
- Using an appropriate response to “don’t know” responses from the child;
- Reading questions verbatim from assessor’s book;
- Reading questions at proper speed;
- Gesturing appropriately in child’s book;
- Using acceptable prompts from assessor’s book;
- Avoiding coaching;
- Keeping the pace of the assessment moving;

- Making eye contact;
- Using neutral praise; and
- Responding to child behaviors.

4. ASSESSOR FEEDBACK AND RECOMMENDATIONS

Assessors were asked to maintain a diary of their experiences during data collection and to participate in debriefings so that the home office staff could learn as much as possible from the field test experience. This chapter presents a summary of the feedback received from assessors and team leaders, as well as the resulting recommendations for revision of the instruments, procedures, and training for the national study.

4.1 Summary of Diaries and Debriefings

Assessors recorded observations about their assessment experiences in field test diaries. Guidelines for completing diaries were reviewed at training, and assessors were advised to record thoughts and comments in their diaries at the end of each assessment day. The completed diaries were returned to the home office once their fieldwork was completed. Diaries were reviewed and summarized by home office staff. Diary summaries were distributed to test developers, the National Center for Education Statistics, and project staff in mid-November.

4.2 Recommendations on Specific Aspects of the Assessment

4.2.1 Assessment Time Frames

The assessors reported that 60 minutes is an appropriate length of time for engaging young children. Many children enjoyed the one-on-one experience and the types of questions asked. However, many assessors noted that the reading passages were too difficult and too numerous for children, especially for children who could not read.

4.2.2 Overall Child Reactions

Assessors reported that the majority of children seemed to enjoy the assessment, liked the one-on-one attention, and approached the experience positively. Science was the section most children seemed to enjoy the most, and they seemed most confident of their answers. Assessors' comments fell into the following four categories: children enjoyed the assessment, children enjoyed the one-on-one attention, children were more alert in the morning than in the afternoon, and children liked science best and reading least.

4.2.3 EFT Cognitive Assessment Items

For the EFT, assessors identified assessment items that they felt were too difficult, poorly presented, did not contain enough guidelines for coding ambiguous responses, or just did not work for most of the children. The following comments were made regarding specific cognitive domains and specific assessment items.

Overall General Comments

Assessors indicated that differences in administration instructions across items made administration of the assessment difficult. They also mentioned that nonreaders had difficulty with the multiple-choice questions in the math and science domains and those children would often ask for the response categories to be repeated. Assessors reported that many children asked about the length of the assessment or asked how much longer the assessment would last.

Mathematics

Assessors indicated that the incorporation of the wooden blocks into the math assessment was not well received by many children. Children were distracted or confused by the blocks and counted the blocks instead of using them to help solve the problem. Assessors recommended that the assessor instructions to present the blocks be written at the top of the assessor's easel so that they are read first rather than last. Assessors also recommended grouping together all of the items that permit the use of the blocks. Assessors reported that the items involving probability and fractions were largely beyond the

kindergarten students that were assessed. Finally, assessors reported that some children were confused by the use of words like “greatest,” “least,” and “fewer.”

Reading

Assessors indicated that differences in administration instructions across items made administration of the assessment more difficult. For example, among similar items, some required sweeping (i.e., having the assessor move his or her hand from left to right under the stimulus picture in order to draw the children’s attention to the picture) while others did not. Additionally, assessors pointed to perceived redundancy in repeating instructions like “Read the sentence and point to the word that goes with it” for many similar items in a row. Assessors noted that kindergarten and first-grade students often had trouble reading the stories written in small print. Additionally, multiple students asked what was meant by the “sound” of the letters.

Science

Assessors noted that some of the science items were hard to understand and had very long answers that overwhelmed some students. Students would usually just select the last response read as their answer because they could not remember the previous three choices. In general, students struggled with a lot of the concepts and content areas presented in the science domain. This was especially true with the kindergarten and first-grade students.

4.2.4 Vision and Hearing Screening

Health technicians reported that the background questions for hearing on the Testing Station Form were sometimes difficult for the children to understand. Additionally, they noted that conditions in the schools were frequently not optimal for screening and that some had lights that were too bright, rooms that were too small, or no separate rooms available. Health technicians said that the equipment was very heavy and that the Pelican cases were unwieldy to manage, particularly in schools with multiple floors and no elevators. Health technicians suggested that an electronic management system, rather than hard copy forms, would help ensure that the data for each child are recorded accurately during the screenings.

4.2.5 Spanish Field Test

Overall General Comments

Assessors reported that most students enjoyed the assessment except for areas where they had no knowledge or could not understand the language. The vocabulary used in the EBRS and SERS was beyond the ability of many of the students assessed, which caused one assessor to recommend the implementation of a skip pattern for these items. Some assessors also recommended combining the warm-up easel and the assessment easel and also putting the "Where's My Teddy?" book at the back of the easel.

SERS

Assessors reported that children had difficulty with many of the items in the SERS. Many children were unable to read the vocabulary words, which discouraged them. Assessors recommended that the vocabulary words should be ordered from least difficult to most difficult. Assessors indicated that they lost children's attention while switching to the "Where's My Teddy" book and then back to the assessment easel. Assessors also had difficulty maintaining the children's attention while repeatedly giving the same instructions (i.e., *Lee esta palabra*). Assessors noted that sweeping on some items and not others also made assessment administration difficult. Lastly, assessors offered suggestions on how they thought the items could be translated more clearly.

EBRS

Assessors reported that some children had difficulty with the vocabulary in the EBRS. Many children were unable to read the vocabulary words, which discouraged them. The assessors also identified a lot of items that were confusing for the children or consistently answered incorrectly. Children often were not familiar with a number of the concepts assessed, including syllables, identifying individual letters, and determining the number of sounds in a word.

4.3 Assessors' Recommendations for Training

The majority of assessors expressed positive comments about the value of training and described the experience as very useful and effective. Role playing with a partner and practice conducting an assessment with a real child were most often cited as the most beneficial parts of training. For the training program that will be conducted prior to fielding the national study, assessors made the following specific recommendations:

- View the assessment materials at the beginning of the training.
- Demonstrate a model role play featuring a complete assessment.
- Present a videotape of a model assessment featuring a real child.

To address these recommendations, a video has been produced that models a full child assessment. This video will be used for all in-person training sessions; it was first used for the school recruiter training in February 2010.

This page intentionally left blank.

5. FIELD TEST ANALYSIS AND DEVELOPMENT OF THE KINDERGARTEN DIRECT COGNITIVE ASSESSMENTS

This chapter describes the analysis conducted by Educational Testing Service (ETS) of the fall 2009 field test direct cognitive assessment data and the recommendations for the development of the national kindergarten direct cognitive assessments.

5.1 Field Test Designs and Item Pools

5.1.1 English Field Test and Item Pool

Cognitive test items in reading, mathematics, and science were administered in the fall 2009 field test—279 unique items in reading, 146 in mathematics, and 171 in science. Items in each subject area were distributed among multiple forms with approximately parallel content and difficulty. Two forms in mathematics and science and four forms in reading per grade combination (kindergarten/first grade, second/third grade) were sorted into 8 booklets per grade combination, totaling 16 booklets in all. Each booklet contained one reading form and one form of either mathematics or science, spiraled such that any of the subject forms appeared in the first or second section of the assessment. The 16 booklets were spiraled among the approximately 3,000 kindergarten, first-, second-, and third-grade test takers participating in the field test. This resulted in approximately 300–800 observations for each test item, dependent upon overlap on other forms within and across grades. Those items appropriate for both kindergarten/first grade and second/third grade were presented on multiple forms and resulted in more observations; others, occurring on only single forms, resulted in fewer observations. Table 5-1 shows the organization.

Approximately 2,600 more respondents in kindergarten through second grade than in third grade participated in the field test. Results were analyzed for all grades combined since the emphasis was on evaluating the performance of the items across a broad range of ability levels and maintaining maximum sample sizes to help stabilize estimates. For issues relating directly to planning for the specific grade testing (e.g., kindergarten round), such as the difficulty of the items, the focus was predominantly on the grade-specific part of the sample only.

Table 5-1. Organization of booklets: 2009

Booklet	Observations	Section 1	Section 2
K/1st Grade Booklet 1	221	K/1st Reading 1	K/1st Mathematics 1
K/1st Grade Booklet 2	223	K/1st Reading 2	K/1st Mathematics 2
K/1st Grade Booklet 3	223	K/1st Reading 3	K/1st Science 1
K/1st Grade Booklet 4	216	K/1st Reading 4	K/1st Science 2
K/1st Grade Booklet 5	217	K/1st Mathematics 1	K/1st Reading 1
K/1st Grade Booklet 6	217	K/1st Mathematics 2	K/1st Reading 2
K/1st Grade Booklet 7	218	K/1st Science 1	K/1st Reading 3
K/1st Grade Booklet 8	212	K/1st Science 2	K/1st Reading 4
2nd/3rd Grade Booklet 1	156	2nd/3rd Reading 1	2nd/3rd Mathematics 1
2nd/3rd Grade Booklet 2	152	2nd/3rd Reading 2	2nd/3rd Mathematics 2
2nd/3rd Grade Booklet 3	155	2nd/3rd Reading 3	2nd/3rd Science 1
2nd/3rd Grade Booklet 4	156	2nd/3rd Reading 4	2nd/3rd Science 2
2nd/3rd Grade Booklet 5	153	2nd/3rd Mathematics 1	2nd/3rd Reading 1
2nd/3rd Grade Booklet 6	151	2nd/3rd Mathematics 2	2nd/3rd Reading 2
2nd/3rd Grade Booklet 7	146	2nd/3rd Science 1	2nd/3rd Reading 3
2nd/3rd Grade Booklet 8	147	2nd/3rd Science 2	2nd/3rd Reading 4

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

A reading form with either a mathematics or science form composed each field test assessment. It is possible that test performance might be improved by an order effect; that is, a test taker may perform better on items administered toward the end of a test, with earlier items serving as practice tasks. Conversely, if a fatigue effect is operating, students may perform better on items administered near the beginning of the assessment, before they become tired. Although order effects were analyzed in the early rounds of the ECLS-K and found to be negligible, the 2009 ECLS-K:2011 field test presented each test form once as the first set of items in the booklet and once as the second. In spite of efforts to reduce test burden through the use of spiraled booklets, the sets of items were still fairly long (55-65 items). This design served the additional purpose of ensuring adequate numbers of responses on each item if some children did not finish both test forms in the booklet.

Reading

Each of the four kindergarten/first-grade reading field test forms had 65 items and 5 to 7 short reading passages. Similarly, the second/third-grade reading field test forms had 70 items and 9 or 10 short reading passages. Several passages and associated items were presented on multiple forms within and across grades. Some of the passages and items were newly developed for the ECLS-K:2011 field test;

others were taken from the operational ECLS-K K–1 and third-grade assessments. Items from the operational ECLS-K assessments were included in the field test in anticipation of inclusion in the ECLS-K:2011 national assessment. The overlap of items across grades and test forms was designed to stabilize item parameter estimates and provide a strong link for selection of test items that will successfully measure gain in successive rounds. The items were a mix of multiple-choice and open-ended formats.[1]

Mathematics

The field test contained 146 unique mathematics items, divided among two forms each for the kindergarten/first-grade and second/third-grade combinations and designed to be approximately parallel within grade with respect to the content and difficulty of the items. Each form appeared in two test booklets, one followed by a reading form and the other with the paired reading form administered first. Some items appeared in multiple forms within or across grades. Others were presented in the ECLS-K operational assessments. As stated above, inclusion of the ECLS-K items was in anticipation of selecting these items for the ECLS-K:2011 national assessment, for linking purposed for longitudinal measures of gain. Both multiple-choice and open-ended items were presented in each form.[1]

Science

Two kindergarten/first-grade and two second/third-grade field test forms each contained 55 items designed to be parallel within content and item difficulty within grade combination. Similar to mathematics, each science form appeared in one test booklet paired with a reading form. Some items appeared on multiple forms, within or across grades. Some items were also previously administered in the ECLS-K operational assessments in anticipation of scaling for longitudinal measurement. Both multiple-choice and open-ended items were presented in the field test.[1]

[1] A table including individual items, reading passages and their associated item sets, and form presentation are available by request as a supplemental document.

5.1.2 Spanish Field Test and Item Pool

Cognitive test items in basic and early reading skills were administered in the Spanish Field Test. The EBRS section, consisting of 28 items in English was followed by the SERS section, consisting of 46 items in Spanish. Both forms were administered (along with two warm-up items) to the approximately 1,000 kindergarten test takers participating in the field test. Results were analyzed for the kindergarten Spanish field test group with data from the kindergarten and first grade English field test and kindergarten and first grade ECLS-K groups. The groups were combined with the emphasis on evaluating the performance of the items in English and Spanish.

English Basic Reading Skills (EBRS)

The 28 items administered were borrowed from the operational ECLS-K or ECLS-B assessments or newly developed for the ECLS-K:2011 and assessed letters and sounds, phonological awareness, and vocabulary. These items were also administered in the English field test on various forms. Items from the operational assessments were included in the field test in anticipation of inclusion in the ECLS-K:2011 national assessment. The items were a mix of multiple-choice and open-ended formats.

Spanish Early Reading Skills (SERS)

Similar to the EBRS, the 46 items administered were borrowed from the operational ECLS-K or ECLS-B assessments or newly developed for the ECLS-K:2011. The skills assessed in the EBRS were also assessed in the SERS, in addition to more complex items measuring sight words, print convention, and ability to locate and recall information. All items and instructions were translated to Spanish. Like the EBRS, the items were a mix of multiple-choice and open-ended formats.

5.2 Field Test Psychometric Analysis

This section describes the psychometric analysis methodology used for evaluating the cognitive field test items. These techniques included item analysis and Item Response Theory (IRT) calibration. Analysis of differential item functioning (DIF) was not carried out because the subgroup

sample sizes for each item were not large enough to support stable estimates. DIF analysis will be conducted on the national test results, and any items found to be unsuitable will be deleted from scoring. The vast majority of items selected for the operational tests had satisfied DIF criteria in previous national administrations. The results of the psychometric analyses of field test items for each cognitive domain are presented in this section.

5.2.1 Methodology

Two different methodologies were used in analyzing psychometric performance of test items: traditional item analysis, which is essentially based on *counts* of right and wrong answers, and IRT analysis, which depends on *patterns* of right and wrong answers and takes into account the differential difficulty of items. Each methodology offers unique perspectives on some aspects of item performance, as well as overlapping views of item difficulty (percent correct versus IRT "b" parameter) and item discrimination (*r*-biserial versus IRT "a" parameter). Item analysis was carried out separately for kindergarteners and first-, second-, and third-graders, so that differences in performance between the grades could be evaluated.

The item analysis tables show, for each item, the number and percentage of students choosing each response option, generally A–D, with the correct option marked with an asterisk (table 5-2). For open-ended items, correct responses are counted under "A" and incorrect responses under "B." The last line for each item shows the mean score on the total set of field test items for students choosing each of the options. The number of students not reaching each item, as well as the number of omits (defined as an unanswered item with a subsequent answered item), is indicated at the left. At the right of the table, the *r*-biserial (adjusted correlation of item performance with total test performance) and P+ (percentage correct) are shown. The point biserial shown on the next line is the simple correlation of the item score with the total, which underestimates the true relationship because the item score (right/wrong) is dichotomous. The *r*-biserial adjusts for this attenuation.

For example, item 52 is an open-ended item, with available options A and B only. The correct response is A, as indicated by the asterisk. The mean score for those selecting the correct response (32.90) is substantially higher than for those selecting incorrectly (24.82). The *r*-biserial is 0.4719 and P+ is 0.1613. Although this is a very hard item—only 16 percent got it right—it may be useful for a high-end

test form because the *r*-biserial shows a strong relationship between performance on this item and on the overall test.

In addition to analysis of each item, summary statistics for each form are presented at the end of each item analysis table. In this example, the alpha coefficient of 0.9251 demonstrates that the set of items has a high level of internal consistency.

Table 5-2. Example of item analysis tables: 2009 field test

					Option						
Category			**Not RCH**	**Omit**	**A**	**B**	**C**	**D**	**Total**		
					A	B	C*	D			
Item	51	N	31	14	49	31	265	44	**434**	R BIS = 0.6021	P+ = 0.6106
		Percent	7.14	3.23	11.29	7.14	61.06	10.14	**100.00**	PT BIS = 0.4736	
4lines		Mean score	19.74	22.50	18.51	20.52	29.52	19.93	**25.74**		
					A*	B					
Item	52	N	31	2	70	331			**434**	R BIS = 0.4719	P+ = 0.1613
		Percent	7.14	0.46	16.13	76.27			**100.00**	PT BIS = 0.3140	
cubeside		Mean score	19.74	20.00	32.90	24.82			**25.74**		

* Correct option.

Number of items analyzed	=	55
Alpha reliability	=	0.9251
Number of cases processed	=	434.0
Minimum score	=	4.0000
Maximum score	=	46.0000
Mean score	=	25.7373
Standard deviation (N)	=	10.0025

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

The unique information provided by the item analysis tables is the performance of the item options: ideally, one should see a mean score for the correct option that is substantially higher than the incorrect option means and no "throwaway" options, that is, options that nearly all students are able to eliminate. Examination of the item analysis tables can identify items that have more than one potentially correct answer (high mean scores for more than one response option) or items that are so difficult that all students appear to be guessing at random (similar mean scores for all options; low *r*-biserial).[2]

The IRT plots, one graph for each test item, show performance of items across the ability range. The horizontal axis, "theta," corresponds to the range of ability of the field test students

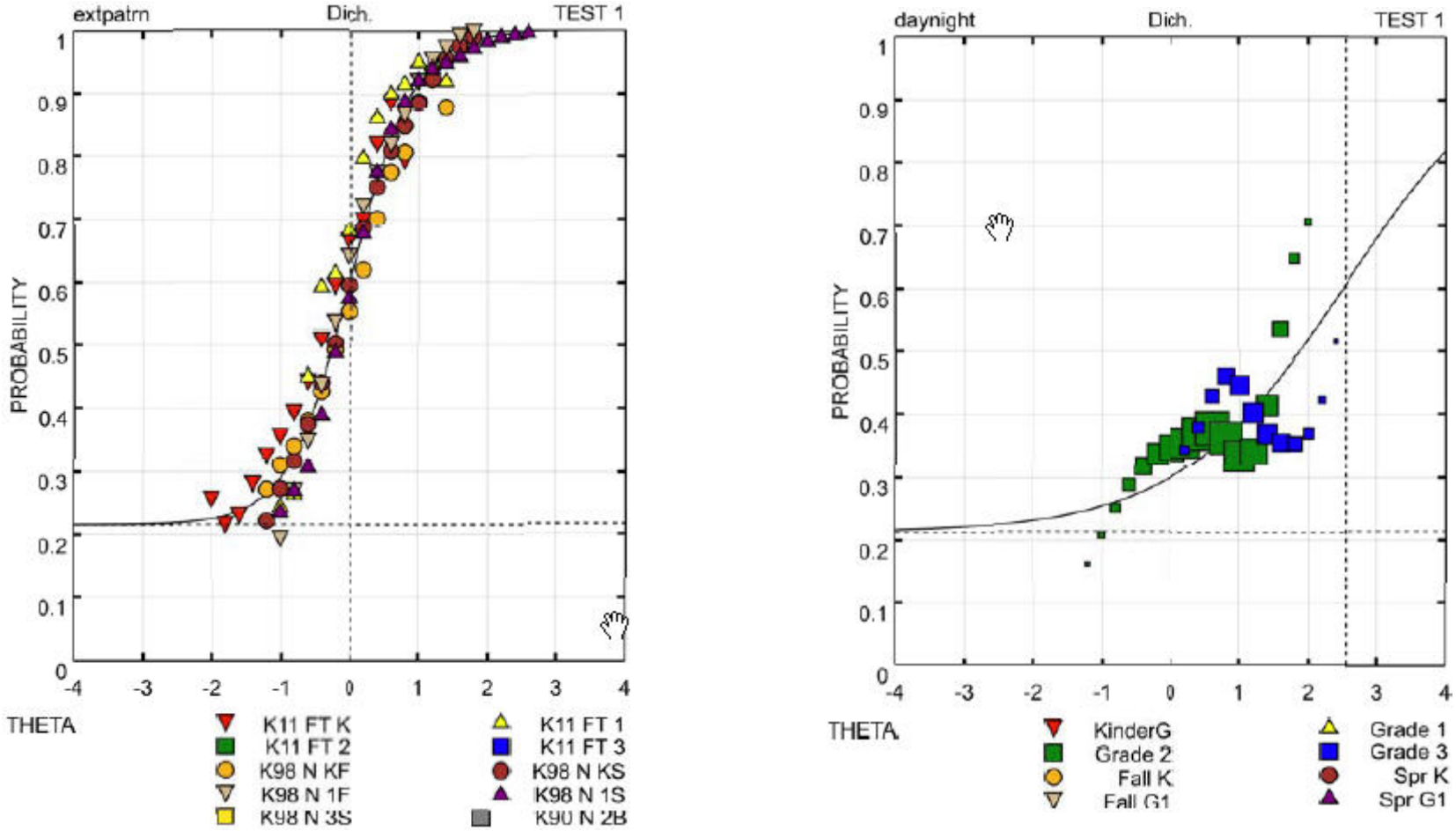
[2] Item analysis results are available by request as a supplemental document.

(figure 5-1). The vertical axis, "probability," indicates the probability of answering the item correctly. The S-shaped curve plotted on the graph shows the fitted model's estimated probability of a correct answer at each point in the ability range, with the horizontal dashed line representing the guessing parameter, that is, the probability of a very low ability test taker getting the item right. The different markers in the graphs in figure 5-1 show the fit of the model to the actual data, separately for different grades from the ECLS-K:2011 field test and the ECLS-K national and bridge rounds. Good items have data that closely fit the curve and a relatively steep slope at the point of inflection of the "S." The IRT "a" parameter, discrimination, is related to the slope at the point of inflection, and is a measure of an item's efficiency in separating test takers whose ability is lower than the corresponding point on the horizontal axis from those of higher ability. The first graph in figure 5-1 shows a successful item, with a steep curve (i.e., success on the item strongly related to overall domain performance) and close fit of data to the model. The second graph is the pictorial representation of a less successful item. Although about 37 percent of children answered correctly, performance on this item is not strongly related to overall science ability: children throughout most of the ability range were about equally likely to answer correctly.[3]

The IRT parameter estimates are less likely than item analysis statistics to be distorted by the omitted items because they are based on the patterns of responses to the items that were answered. In analyzing the field test results, IRT estimates of difficulty and discrimination were given more weight than the analogous statistics, P+ and *r*-biserial, from traditional item analysis. However, the traditional item analysis provides information on individual item options that cannot be observed in the IRT results—for example, items that have more than one potentially correct answer, response options that may be confusing or misleading, or options that are so implausible that they were chosen by very few children.

[3] IRT plots are available by request as a supplemental document.

Figure 5-1. Examples of IRT plots, ECLS-K:2011 fall 2009 field test



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

5.2.2 Analysis Results

Below is a discussion of the adequacy of the field test item pool, directly followed by sections that summarize analysis findings for the reading, mathematics, science, EBRS, and SERS field test items.

Adequacy of the English and Spanish Field Test Item Pools

The items field tested in fall 2009 were drawn from several sources: operational items from the ECLS-K rounds 1–4 (kindergarten and first grade),[4] ECLS-K round 5 (third grade),[5] and the ECLS-B preschool and kindergarten rounds.[6] Each of these sources has a large number of additional items available for use in ECLS-K:2011 operational assessments. In addition, approximately 118 reading, 80 mathematics, and 142 science items were newly developed for the field test to enable analyses of changes in education policy, pedagogy, early childhood research, and society since the earlier studies, as well as the flexibility to anticipate new policies and research yet to emerge and to incorporate elements into the study that are designed to address them.

- A field test of cognitive items generally has the following objectives:
- Evaluate item quality and identify flaws in wording or response options for possible revision.
- Ascertain the range of achievement likely to be encountered in the sample of students who will later take the operational test.
- Calibrate the field test item difficulties on the same scale of student achievement, so that items of appropriate difficulty may be selected for the final forms.
- For the Spanish field test, evaluate performance of items in both English and Spanish, and for English- and/or Spanish-speakers.

[4] U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Kindergarten and First-Grade Operational Assessments, fall 1998 through spring 2000.

[5] U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Third-Grade Operational Assessment, spring 2002.

[6] U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Preschool and Kindergarten Operational Assessments, fall 2005 through spring 2007.

By design, approximately 80 percent of the items field tested has been used before, either in the ECLS-B or ECLS-K K–1 or third-grade large-scale operational tests, so item-quality concerns have already been addressed for these items. In both the English and Spanish field test analyses, particular attention to item-quality issues was paid to data from the items newly developed for the ECLS-K:2011. In the Spanish field test analysis, additional analyses were performed examining the effect of language on item function. As expected, the field test results did not show substantial numbers of legacy items that needed to be discarded or revised. In fact, for items that will be used as common items to equate to the ECLS-K scale, it is important that the items *not* be revised. The field test results did, however, show that some newly developed items required revisions or should be discarded, while others functioned well at the assessed grade levels.

In order to measure each child's status accurately, it is important that each child receive a set of test items that is appropriate to his or her skill level. Selection of potential items brings together two sets of information: the difficulty parameters for each of the items in the pool and the range of ability expected in each round. 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. IRT calibration of the English field test item data was carried out for each subject area by pooling the following datasets together:

- ECLS-K:2011 fall kindergarten field test (approximately 890 cases)
- ECLS-K:2011 fall first-grade field test (approximately 850 cases)
- ECLS-K:2011 fall second-grade field test (approximately 800 cases)
- ECLS-K:2011 fall third-grade field test (approximately 400 cases)
- ECLS-K fall kindergarten national (approximately 18,000 cases)
- ECLS-K spring kindergarten national (approximately 19,000 cases)
- ECLS-K fall first-grade national (data collected only for a subsample of about 5,000)
- ECLS-K spring first-grade national (approximately 16,000 cases)
- ECLS-K spring second-grade bridge sample (approximately 900 cases)
- ECLS-K spring third-grade national (approximately 14,000 cases)

The ECLS-K had no separate science assessment in the kindergarten and first-grade rounds; science items from the K–1 general knowledge test were selected and pooled with the other datasets.

The Spanish field test data (approximately 1000 cases) was pooled with only the kindergarten and first grade rounds of the ECLS-K:2011 field test and ECLS-K national rounds for a separate IRT calibration focusing on the effects of language of administration at the school-entry level.

The overlapping items shared by two or more datasets serve as anchors, so that parameters for items and samples from different assessments are all on a common scale. Data from the ECLS-K:2011 field test supply a link between the newly developed field test items and the ECLS-K assessments, enabling them to be put on the common scale necessary for direct comparisons. The large samples from the ECLS-K data collections also serve to stabilize parameter estimates that would be unreliable if only the relatively small sample of the fall 2009 field test were available.

Pooling the datasets together also provides estimated values for the mean ability levels for each group on the same scale. Although the datasets are pooled, the samples are identified individually so that the ability range of each group can be obtained. The mean and standard deviation of the ability levels for each of the groups above were calculated from the pooled sample. Therefore, an estimated ability range for the target administrations (e.g., fall and spring kindergarten) can be determined.

The pool of items available for assembly of the national test forms is not limited to the items in the 2009 field tests. Using the methodology described the difficulty parameters were calibrated in a common metric for all of the items used in all of the datasets, regardless of whether they appeared in the 2009 field test forms. Thus, virtually *all* items in the source tests can be considered part of the item pool for the purpose of test assembly.

5.2.2.1 English Field Test

Reading

Overall, the field test items for reading performed well, as expected. From the item analysis, the majority of *r*-biserials were well above the desired value of 0.3. Of the 279 unique items administered, 31 exhibited *r*-biserials lower than ideal. Of that 31, 12 were very easy items (P+ values greater than 0.9),

and another 12 were quite difficult (P+ values less than chance for the multiple-choice items, or less than .1 for open-ended items). The remaining seven items exhibited low *r*-biserials because children who chose the correct response had an average score that was just slightly higher than those who chose an incorrect response. Items such as this were not included in the proposed forms for the national assessment in order to remove any ambiguity in selection of a single correct response. All of the other items showed the expected trends in response selection; the correct response was selected by groups of students who have higher total scores. The alpha reliabilities for each of the eight test forms were approximately 0.95, well above the targeted minimum of 0.8.

Review of the IRT plots showed good fit of item data with the estimated parameters. Although the fit was good for most of the items, the discrimination was not necessarily so. This generally occurs with items that are either relatively easy or hard. In selecting items for the national forms, items with poor fit and discrimination are avoided.

Mathematics

As with reading, the field test items for mathematics performed well, as expected. From the item analysis, the majority of *r*-biserials were above the desired value of 0.3. Of the 146 unique mathematics items administered in the field test assessment, 10 had *r*-biserials lower than ideal, all either because the item was too easy (N=6) or too difficult (N=4). The alpha reliabilities for each of the four test forms were approximately 0.93.

Review of the IRT plots showed good fit of item data with the estimated parameters for most of the items. For those items in which the fit was poor, the item was generally too hard or too easy. Although the fit was good for most of the items, the discrimination was not necessarily so. Similarly, this generally occurs with items that are relatively hard. In selecting items for the national forms, items with poor fit and discrimination are avoided.

Science

Unlike the reading and mathematics items, fewer of the field test items for science performed well. Of the 171 unique science items, 72 exhibited *r*-biserials lower than the desired value of

0.3. Upon analysis of the field test data, it became apparent that assessing children in science in fall kindergarten was not necessarily appropriate for the ECLS-K:2011, as illustrated by the item characteristic curves for a majority of the items. For items administered at the other grade levels, the majority functioned well and thus provide an adequate pool from which to build test forms at these levels, as well as justification to assess at these levels in the national administration. At the kindergarten level, however, more often than not items showed inconsistent behavior in relation to what was expected and to the other grade levels. For example, in some cases the item was just too difficult for kindergarten. Another set of items showed that the kindergarten data were not consistent with data on the same item for first-graders. Still another set showed that the kindergarten children were guessing across the ability range. These patterns of performance were consistent across many of the items assessed in kindergarten. Thus, as discussed in detail below, a short kindergarten science assessment was developed for the spring data collection round only.

The alpha reliabilities for each of the four test forms were approximately 0.83–0.91. These are lower than those for the reading and mathematics forms, but are not unexpected, since the variability in item content is greater for science.

5.2.2.2 Spanish Field Test

English Basic Reading Skills

The majority of field test items for the EBRS performed well in the Spanish field test. From the item analysis, all items except one exhibited *r*-biserials well above the desired value of 0.3. The one item with an *r*-biserial slightly below ideal was quite difficult with a P+ value near .1. This item was excluded from the proposed form for the national assessment. The alpha reliability of the EBRS items in the Spanish field test was 0.87, above the targeted minimum of 0.8. Review of the IRT plots showed good fit of item data with the estimated parameters.

Spanish Early Reading Skills

Overall, the field test items for the SERS performed well. From the item analysis, the majority of *r*-biserials were well above the desired value of 0.3. One difficult vocabulary item and three

quite difficult items reading items at the end of the assessment exhibited *r*-biserials lower than ideal. The alpha reliabilities for the SERS form was approximately 0.89, above the targeted minimum of 0.8. As with the EBRS, review of the IRT plots showed good fit of item data with the estimated parameters.

5.3 Design of the Kindergarten Tests

This section describes the design of the national assessment forms. Numerous competing objectives were taken into account in selecting reading passages and reading, mathematics, science, EBRS, and SERS items for the proposed forms, including the following:

- **Difficulty:** Matching the difficulty of the test questions to the expected range of ability that will be found in the national administrations; choosing routing questions and second-stage forms of appropriate difficulty; avoiding floor and ceiling effects (sets of items that are *all* too hard or too easy for some of the children taking them).
- **Test specifications:** Matching as closely as possible the target percentages of content categories and of old and new test items.
- **Psychometric characteristics:** Selecting items that do a good job of discriminating among achievement levels.
- **Linking:** Having sufficient overlap of items shared among forms and across rounds so that a stable scale can be established for measuring status and gain, as well as having an adequate number of items carried over from the ECLS-K to permit cross-cohort comparisons.
- **Proficiency levels:** Retaining items from the ECLS-K assessments that are necessary for measuring status with respect to established proficiency levels.
- **Assessor feedback:** Incorporating recommendations made by the field staff based on their observations of item functioning.
- **Measurement of gain:** Evaluating whether continued improvement in performance on items could be expected in the years beyond those being assessed.
- **Time limits:** Making efficient use of testing time, both to minimize burden on test takers and schools and for budgetary reasons.

Ability estimates of the kindergarten national sample.[7] IRT ability estimates were used to define target difficulty ranges for different forms of the kindergarten test. The ability (theta) estimates for the ECLS-K:2011 fall kindergarten field test and the ECLS-K fall and spring kindergarten sample, from the pooled analysis above, are used as estimates of the range of abilities that can be expected in the ECLS-K:2011 national fall and spring kindergarten samples.

The analysis results showed discrepancies in estimated mean ability levels between the ECLS-K:2011 field test and ECLS-K national samples. The ECLS-K national mean at fall kindergarten is about 0.6–0.7 of a standard deviation (based on either the ECLS-K:2011 or ECLS-K standard deviation) below the mean of the fall 2009 kindergarten field test sample in reading and mathematics, and about 0.2–0.3 of a standard deviation below in science. The gap in mean ability between the ECLS-K national fall first-grade mean and the ECLS-K:2011 fall field test first-grade mean is similar. Several factors may contribute to these discrepancies. First, a field test sample is necessarily small relative to the population and may not be perfectly representative. The range of abilities in the field test sample may be somewhat attenuated, probably at the low end. It is possible that the ECLS-K:2011 field test sample included a disproportionate number of children with higher than average abilities. Another possible factor could be real changes in the kindergarten population in the interval between 1998 and 2009 with respect to demographics and/or prior exposure to early learning. Expansion of preschool programs serving disadvantaged children could result in kindergarten entrants in 2009 being, on average, better prepared for school than those who entered kindergarten in 1998. Without knowing the explanation for the discrepancy with certainty, the range of difficulty of the proposed test forms was targeted to be suitable for the whole range of ability levels found from the low end of the ECLS-K distribution to the upper end of the ECLS-K:2011 field test range. The targeted means and ability ranges for each domain are discussed below.

Continuity of the ECLS-K proficiency levels. In the early rounds of the ECLS-K, proficiency levels consisting of clusters of test items were identified as a means of analyzing mastery of developmental milestones. Ideally, an analysis of longitudinal growth should take into account not only the number of scale score points gained from time 1 to time 2, but also where on the continuum of achievement the gains took place. The proficiency probability scores in the ECLS-K facilitate meaningful analysis of relationships between gains and variables such as school processes, demographics, and home

[7] Estimates for the first- and second-grade samples and form designs will be followed as an addendum to this report. Results from the kindergarten national data collection analysis will inform the estimates and form designs at these grade levels. Since the analysis will not be completed prior to delivery of this report, the addendum will be delivered separately.

environment measures. By round 5, third grade, eight proficiency levels had been defined for the reading and seven for the mathematics assessments, and analysis of the data confirmed that measured growth tended to follow the hypothesized hierarchical model. No proficiency levels were developed for the science assessment, because science curriculum is more diverse and cannot be assumed to follow a hierarchical sequence.

Five proficiency levels were identified in each subject, reading and mathematics, in the ECLS-K kindergarten through first-grade assessment instruments. In designing each subsequent assessment, performance on the most recent round, along with field test results for the next round, were taken into account in determining the appropriate amount of overlap of proficiency levels from one assessment to the next. Tables 5-3 and 5-4 list the proficiency levels in ECLS-K assessments through round 5, third grade.

Table 5-3. ECLS-K proficiency levels in reading, through third grade

Proficiency level	Description
1	Letter recognition
2	Beginning sounds
3	Ending sounds
4	Sight words
5	Words in context
6	Literal inference
7	Extrapolation
8	Evaluation

SOURCE: Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), *Psychometric Report for the Third Grade* (NCES 2005–062), chapter 4 and appendixes A, B, and C.

Table 5-4. ECLS-K proficiency levels in mathematics, through third grade

Proficiency level	Description
1	Number and shape
2	Relative size
3	Ordinality, sequence
4	Addition/subtraction
5	Multiplication/division
6	Place value
7	Rate and measurement

SOURCE: Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), *Psychometric Report for the Third Grade* (NCES 2005–062), chapter 4 and appendixes A, B, and C.

The 2009 field test forms include some but not all of the items marking the ECLS-K proficiency levels in reading and mathematics. As noted above, even items that were not field tested can

be selected for operational forms because the pooling of datasets results in calibration on a common scale. After each ECLS-K:2011 national round of data collection, test results will be analyzed in an attempt to determine if the ECLS-K proficiency level structure, definitions, and procedures are appropriate for the ECLS-K:2011 data. An alternative option would be to develop a new set of proficiency levels, as was done for the ECLS-B preschool and kindergarten assessments.

5.3.1 Reading and EBRS

As discussed above, the range of difficulty of the proposed test forms targeted from the low end of the ECLS-K distribution to the upper end of the ECLS-K:2011 field test sample. With this assumption, the range from roughly two standard deviations below the fall ECLS-K mean (-2.27), to two standard deviations above the estimated[8] spring ECLS-K:2011 mean (+1.29), should include at least 95 percent of the ECLS-K:2011 national kindergarten sample, even if the discrepancy between the ECLS-K:2011 field test and the ECLS-K national results remains unexplained. Table 5-5 lists the estimated means and standard deviations of theta for kindergarten.

Table 5-5. Estimated means and standard deviations of theta for kindergarten: Reading

Sample	Mean theta	Standard deviation of theta
Fall kindergarten – ECLS-K:2011 field test	-0.60	0.63
Fall kindergarten – ECLS-K national	-1.03	0.62
Spring kindergarten – ECLS-K:2011 field test (estimated)	+0.03	0.63
Spring kindergarten – ECLS-K national	-0.39	0.59

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

The range of -2.27 to +1.29 defines not only the ability range of the children, but also the corresponding difficulty parameters of the items required for the assessment. Of course, the parameter estimates are provisional and will be recalibrated for the national administration. There are many factors that may contribute to slight changes in the parameter estimates: the assortment of items selected for the national test and the order in which they will be given, the number and location of practice items, discontinue rules, real differences between the field test and historical samples compared with the

[8] The spring kindergarten ECLS-K:2011 mean theta was calculated by assuming the growth from fall kindergarten to spring kindergarten to fall first grade is the same as what was observed in the ECLS-K. That is, the children in the ECLS-K sample gained a little more than one standard deviation from fall to spring kindergarten, and nearly another half standard deviation by fall first grade. This is consistent with what was observed in the ECLS-K:2011 field test sample, which gained a little under 1.5 standard deviations from fall to fall, kindergarten to first grade.

population of kindergarten entrants in 2010, etc. As a precaution against encountering floor and ceiling effects in the national assessment, it was recommended to extend the difficulty range of the items at both the low and high ends. Thus, some items with b parameters below -2.27 at the low end and above +1.29 at the high end have been included in the national forms.

English Basic Reading Skills and Routing

The design of the ECLS-K:2011 reading assessment incorporates a component to measure English basic reading skills (EBRS) for all children, regardless of home language. It was anticipated that the data from the English (non-ELL) and Spanish (ELL) samples would show differences across the ability/difficulty ranges. This did not prove to be the case: item data from the ELL group tracked data from the non-ELL group, with the model fitting data from both samples well. With this result, a subset of items from the SFT was selected for the national administration, referred to as the EBRS.

The goal of the EBRS is to measure basic reading skills in English for all children, regardless of home language; thus, the EBRS items will be administered to all children in the national sample. The EBRS items will also serve as part of the reading routing test, with additional routing items administered to children who perform well on the EBRS set. The items on the EBRS are relatively easy and would not be adequate to distinguish differences between children at middle- and high-ability levels, and thus the need for a second, more difficult, set of routing items.

Routing and Second-Stage Form Design

The psychometric characteristics of the items were reviewed, and any items that were unsatisfactory with respect to the quality criteria described above were deleted. The items were then sorted according to content category and presentation. For example “basic skills” items were presented in several different formats. The difficulty statistics were reviewed for the items within each content/presentation type, and each set was classified suitable for the routing test (either Part 1 [EBRS] or Part 2), the low, middle, or high form, according to the difficulty of the majority of the items in each set. The different presentations of the same content were compared, and where there was redundancy, the item sets with the strongest psychometric characteristics were selected. In general, the types were ordered in increasing order of average difficulty (although most had a spread of difficulty within types), but other

factors were also taken into consideration, such as grouping items by format to minimize changes in task instructions.

The distributions of thetas described above define the range of abilities to be targeted by the test forms. The IRT difficulty parameters for the pool of available items are calibrated on the same scale as the abilities. Thus, the process of choosing test items relies on matching the difficulty of the items to the abilities of the test takers. To optimize the measurement accuracy of the tests, the selected items should be approximately equally spaced along the ability/difficulty scale.

Table 5-6 shows the estimated ability ranges for low-, middle-, and high-level groups for fall and spring kindergarten. Also shown in the table are the number of items in the peak range on each form and which form will likely be administered to each ability group. For example, middle-ability children in spring kindergarten are expected to have ability estimates between -0.98 and +0.66. They would receive both sets of routing items followed by form B. These sets of items would include 31 (9+15+7) items whose difficulty matches their anticipated ability level. (Note that not all items fall within the peak range in the second-stage forms. Items outside the peak range are a result of the intentional addition of items to extend difficulties beyond the peak range and items to provide overlap between forms needed to support development of a common score scale.) For each range in the table, the low end of the range was computed using the mean and standard deviation of the lower scoring sample (ECLS-K national), while the high end of the range was based on the mean and standard deviation of the ECLS-K:2011 field test sample.

Table 5-6. Peak and full difficulty ranges, routing plus second stage: Reading

	Fall kindergarten			Spring kindergarten		
Item	**Low level (-2sd to mean)**	**Middle level (+/- 1sd)**	**High level (mean to +2sd)**	**Low level (-2sd to mean)**	**Middle level (+/- 1sd)**	**High level (mean to +2sd)**
Estimated peak ability range (95% of sample)	-2.27 to -0.60	-1.65 to -0.03	-1.03 to +0.66	-1.57 to +0.03	-0.98 to +0.66	-0.39 to +1.29
Number of items in peak difficulty range:						
Part 1 Routing (EBRS)	13*	15*	10*	15*	9*	2*
Part 2 Routing	0	8*	20*	8*	20*	20*
Form A	11*	9	2	7	2	0
Form B	8	9*	8	7*	8*	5
Form C	1	1	11*	1	11	19*
Form by design (designated by *)	R1+A	R1+R2+B	R1+R2+C	R1+R2+B	R1+R2+B	R1+R2+C

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

Other Factors Considered in Form Design

Other factors were also considered in assembling the test forms: 1) the framework specifications for content categories of items, 2) selection of items for which continued growth is expected from fall kindergarten to spring kindergarten to first grade to support measurement of longitudinal gains, and 3) the overlap of items with the ECLS-K to support cross-cohort comparisons.

Framework

The framework design for the ECLS-K:2011 was derived from the ECLS-K test specifications. The leftmost column of table 5-7 lists the targeted content categories:

- BS: Basic skills
- VOC: Vocabulary
- LOC: Locate/recall
- INT: Integrate/interpret
- CRIT: Critique/evaluate

Table 5-7. Framework targets and items by content area: Reading

		Including EBRS items			Excluding EBRS items		
Content area	Targeted percent of items	Targeted number of items	Proposed number of items	Proposed percent of items	Targeted number of items	Proposed number of items	Proposed percent of items
BS	50	42	53	64	32	35	56
VOC	15	12	11	13	9	9	14
LOC	20	17	14	17	13	14	22
INT	10	8	3	4	6	3	5
CRIT	5	4	2	2	3	2	3
TOTAL	100	83	83	100	63	63	100

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

The next column shows the content percentages targeted by the framework. Target and proposed number and percentage of items are listed for the item pool including the EBRS items in the

next three columns and excluding the EBRS items in the last three. Comparisons were done both ways since the EBRS items are predominantly from the basic skills category. The requirement to include the full set of 20 EBRS items to measure basic reading skills in English skewed the content category percentages to include more basic skills items. For this reason, table 5-7 shows targeted percentage of items by content category compared with the proposed percentages both with and without the required EBRS items.

Design of the reading assessment is somewhat different from the other domains since the items associated with reading passages must be selected in sets rather than individually. A reading passage may be desirable if it has one or more associated items in a content category that is hard to fill in the first years of school, such as INT or CRIT. But the set may also include some basic skills items, which are already overrepresented. This presents an additional challenge in selecting reading items, related to time constraints, which is not present for the other assessment components. The investment of time required for reading a passage before answering the questions means that only a few passages can be included and suggests that as many associated items as possible should be selected for each. But this can impact the balance of framework proportions on the test as a whole.

On the high second-stage form, passages and their associated item sets are included to target the anticipated ability levels and content categories designated by the framework. The passage sets selected for the high form maximized the number of integrate/interpret and critique/evaluate items in the appropriate difficulty range for fall and spring kindergarten. Even with this maximization, however, both categories fall short of the targets. The available item pools did not include items in these categories that performed well for children who do not yet have the skills required to read and understand text. Items in the locate/recall category fall close to the targeted proportion, while those from vocabulary are slightly lower than targeted.

The percentage of items from the basic skills category is slightly higher than targeted, and even more so when the EBRS items are included in the counts. The ECLS-K:2011 incorporates more of a variety of basic skills item types than the original ECLS-K, based on the framework. Table 5-8 lists the different subcategories of basic skills items in the proposed item pool, and the number and percent of these items. (The numbers in table 5-8 reflect the proposed pool *with* the EBRS items included.) The phonemic awareness, phonemic substitution, segmentation, blending, and rhyming item types (a total of 14 items) were not included in the original ECLS-K assessments.

Table 5-8. Subcategories of basic skills items in proposed pool: Reading

Basic skills subcategory	Number of items in proposed pool	Percent of items in proposed pool
Letter recognition	6	7.2
Letter sounds	4	4.8
Beginning sounds	5	6.0
Ending sounds	4	4.8
Phonemic awareness	6	7.2
Phonemic substitution	2	2.4
Segmentation	2	2.4
Blending	2	2.4
Rhyming	2	2.4
Sight words	12	14.5
Print convention	8	9.6
Syllables[1]	0	0
Total	53	64

[1] Items asking the child to respond with the number of syllables in a word were too difficult, did not function well, and received assessor feedback that the items were confusing to kindergarteners. They were excluded from the national assessment.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

As stated above, inclusion of passage item sets and phonemic item subcategories in basic skills, in addition to the requirement that all children are administered the EBRS item set, resulted in the proposed deviation from the target framework categories.

Item Overlap

The development of a vertical scale that must span kindergarten to second grade and beyond and have optimal measurement properties throughout the achievement range calls for multiple test forms that vary in their difficulty. The forms are tailored for individuals within the targeted ability levels. However, the overall item pool at each grade should reflect core curriculum elements for that particular grade level. At the same time, there must be overlapping items shared by forms within a round, as well as across rounds. These linking items tie the vertical scale together both across forms within a round and across rounds. In general, at least 20–30 percent of the items should overlap for adequate scaling.

Tables 5-9 and 5-10 show the overlap of items across forms within the proposed kindergarten reading assessment and the overlap with the legacy assessments. The 20 Routing Part 1 (EBRS) items, taken by all children, and the 20 Routing Part 2 items, taken by a subset of children, also serve as common items for linking across all forms. Based on the rule of thumb listed above, there are

ample numbers of items across forms and across assessments to create both horizontal (i.e., within-round, and cross-cohort) and vertical (i.e., longitudinal) scales.

As defined in the original specifications, the ECLS-K:2011 targeted 80 percent of items from the original ECLS-K and/or ECLS-B assessments. As a consequence of introducing a variety of new phonemic awareness items in the basic skills category, a shortfall in legacy items resulted. It was recommended to accept this discrepancy, since the total number of items overlapping is still more than adequate for cross-cohort comparisons with the ECLS-K.

Table 5-9. Number of items overlapping across forms: Reading

Form	Number of unique overlapping items
Forms A and B	7
Forms B and C (Routing Part 2 items)	20
Forms A, B, and C (including Routing Part 1 [EBRS] items)	21

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

Table 5-10. Number of items in proposed kindergarten assessment overlapping with the ECLS-K: Reading

Assessment	Number of unique overlapping items	Percent of unique overlapping items
ECLS-K K-1, ECLS-K 3rd grade, and/or ECLS-B national	63	76

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

5.3.2 Mathematics

As described above, the average ability estimated for the kindergarten field test participants was about 0.6 to 0.7 of a standard deviation higher than that of the ECLS-K fall kindergarten national sample. With no way of knowing whether this discrepancy reflects real population differences or is an artifact of the restricted field test sample, it was necessary to design the assessment in a way that would provide for accurate measurement in either circumstance. Ability levels roughly two standard deviations below the fall ECLS-K theta mean in mathematics (-2.25), through two standard deviations above the estimated[9] spring ECLS-K:2011 mean (+1.16), should include about 95 percent of the ECLS-K:2011

[9] The spring kindergarten ECLS-K:2011 mean theta was calculated by assuming the growth from fall kindergarten to spring kindergarten to fall 1st grade is the same as what was observed in the ECLS-K. That is, the children in the ECLS-K sample gained about one standard deviation

national kindergarten sample. Table 5-11 lists the estimated means and standard deviations for fall and spring kindergarten in mathematics.

Table 5-11. Estimated means and standard deviations of theta for kindergarten: Mathematics

Sample	Mean theta	Standard deviation of theta
Fall kindergarten – ECLS-K:2011 field test	-0.60	0.53
Fall kindergarten – ECLS-K national	-0.99	0.63
Spring kindergarten – ECLS-K:2011 field test (estimated)	-0.10	0.53
Spring kindergarten – ECLS-K national	-0.40	0.59

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

As in the design of the reading forms, the difficulty range of the selected items at both the low and high ends was extended to avoid floor and ceiling effects. Some items with difficulty parameters below -2.25 at the low end and above +1.16 at the high end are included. Table 5-12 shows the estimated ability ranges for low-, middle- and high-level groups for fall and spring kindergarten. Also shown are the number of items in the peak range on each form and which form is designed to be administered to each level group. For fall kindergarten, most children are expected to be administered Form A or B, and in spring, most children to be administered Form B or C.

Table 5-12. Peak and full difficulty ranges, routing plus second stage: Mathematics

	Fall kindergarten			Spring kindergarten		
Item	Low level (-2sd to mean)	Middle level (+/- 1sd)	High level (mean to +2sd)	Low level (-2sd to mean)	Middle level (+/- 1sd)	High level (mean to +2sd)
Estimated peak ability range (95% of sample)	-2.25 to -1.13	-1.62 to -0.07	-0.36 to +0.47	-1.58 to -0.44	-0.99 to +0.62	+0.19 to +1.16
Number of items in peak difficulty range:						
Routing	4	12	7	9	13	2
Form A	10*	9	0	8	3	0
Form B	4	19*	6*	10*	18*	2
Form C	0	5	6	3	14	18*
Form by design (designated by *)	A	B	B	B	B	C

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

from fall to spring kindergarten, and another half standard deviation by fall first grade. This is consistent with what was observed in the ECLS-K:2011 field test sample, which gained a little over 1.5 standard deviations from fall to fall, kindergarten to first grade.

Other Factors Considered in Form Design

As with the reading assessment design, the following additional factors were also considered in assembling the test forms.

Framework

The framework design for the ECLS-K:2011 was derived from the ECLS-K test specifications. The leftmost column of table 5-13 lists the targeted content categories:

- NSPO: Number sense, properties, and operations
- MEAS: Measurement
- GSS: Geometry and spatial sense
- DSP: Data analysis, statistics, and probability
- PAF: Patterns, algebra, and functions

Table 5-13. Framework targets and items by content area: Mathematics

Content area	Targeted percent of items	Proposed number of items	Proposed percent of items
NSPO	75	57	76
MEAS	5	2	3
GSS	3	2	3
DSP	8	6	8
PAF	9	8	11
TOTAL	100	75	100

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

The proposed percentages of items hit the targeted percentages in the GSS and DSP content areas. The shortfall in the MEAS category is due to the lack of potentially successful measurement items in the pool consisting of the newly developed items for the field test, the legacy items fielded, and the legacy items not fielded. That is, all other measurement items were outside of the difficulty range suitable for kindergarten or the psychometric characteristics were poor. Conversely, the mathematics item pool included more items from the PAF and NSPO categories than targeted. The distribution of item

difficulties required that these additional PAF and NSPO items be included to ensure accurate measurement at various intervals across the distribution.

Item Overlap

Tables 5-14 and 5-15 show the overlap of items across forms within the proposed kindergarten assessment and the overlap with the legacy assessments. The 18 routing items, taken by all children, also serve as common items for linking across all forms. There are adequate numbers of items across forms and across assessments to create both horizontal (i.e., within-round and cross-cohort) and vertical (i.e., longitudinal) scales. As defined in the original specifications, the ECLS-K:2011 targeted 80 percent of items used in the original ECLS-K and/or ECLS-B assessments.

Table 5-14. Number of items overlapping across forms: Mathematics

Form	Number of unique overlapping items
Forms A and B	4
Forms B and C	6
Forms A, B, and C (including routing items)	19

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

Table 5-15. Number of items in proposed kindergarten assessment overlapping with the ECLS-K: Mathematics

Assessment	Number of unique overlapping items	Percent of unique overlapping items
ECLS-K K-1, ECLS-K 3rd Grade, and/or ECLS-B national	60	80

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

5.3.3 Science

The science items were reviewed based on *all* of the available information. It became apparent that assessing children in science in fall kindergarten was not appropriate for the ECLS-K:2011, as illustrated by the item characteristic curves for a majority of the items. For first grade and beyond, the majority of items functioned well. Analysis results show an adequate pool from which to build test forms

at these grades, as well as an indication that a science assessment would be successful in the national administration. At the kindergarten level, however, more often than not, items showed inconsistent behavior in relation to what was expected and to the other grade levels. For example, in some cases the item was just too difficult for kindergarten. Another set of items showed that the kindergarten data were not consistent with data on the same item for first-graders. Still another set showed that the kindergarten children are guessing across the ability range. These flaws in measurement properties affected many or most of the items assessed in kindergarten.

Based on the field test analysis findings, assessing children in fall kindergarten does not seem worthwhile. Although some items did function well in kindergarten, the number of items was quite limited and not enough to justify recommending a fall kindergarten assessment. Moreover, the items that did have acceptable performance were predominantly life science items, with only a few successful physical science and almost no earth science items. This would make it impossible to select a set of items for a full-scale kindergarten science assessment consistent with the test framework. For the other grade levels, there were adequate numbers of items in each category that functioned well. Thus, it is recommended to develop the proposed two-stage assessments for first and second grades.

With this recommendation, however, the collection of science data would begin on only a subsample of the ECLS-K:2011 in fall first grade. Therefore, it is also recommended to administer a limited, 20-item, single-stage test in spring kindergarten. This smaller assessment would permit measurement on the entire sample on a limited set of items appropriate for spring kindergarten and would calibrate with subsequent rounds of science data collection, thus providing an early data point in science.

With this assumption, data roughly two standard deviations below the spring science ECLS-K theta mean (-1.70), and two standard deviations above the estimated[10] spring ECLS-K:2011 mean (+1.35), should include about 95 percent of the ECLS-K:2011 national spring kindergarten sample. Table 5-16 lists the estimated means and standard deviations for spring kindergarten.[11]

[10] The spring kindergarten ECLS-K:2011 mean theta was calculated by assuming the growth from fall kindergarten to spring kindergarten to fall first grade is the same as what was observed in the ECLS-K. That is, the children in the ECLS-K sample gained about half of a standard deviation from fall to spring kindergarten, and another third of a standard deviation by fall first grade. This is consistent with what was observed in the ECLS-K:2011 field test sample, which gained about three-fourths of a standard deviation from fall to fall, kindergarten to first grade.

[11] The ECLS-K did not have a separate science assessment in the kindergarten and first-grade rounds, only the general knowledge test that was a mix of science and social studies questions. The statistics presented here for the ECLS-K are based on the science items from the general knowledge test.

Table 5-16. Estimated means and standard deviations of theta for spring kindergarten: Science

Sample	Mean theta	Standard deviation of theta
Spring kindergarten – ECLS-K:2011 field test (estimated)	+0.05	0.65
Spring kindergarten – ECLS-K national	-0.04	0.83

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

As with the reading and mathematics form design, items are included with difficulty parameters below -1.70 at the low end, and above +1.35 at the high end to avoid floor and ceiling effects.

Other Factors Considered in Form Design

As with the reading and mathematics form designs, additional factors were also considered in assembling the test forms.

Framework

The framework design for the ECLS-K:2011 was derived from the ECLS-K test specifications. The leftmost column of table 5-17 lists the targeted content categories:

- Scientific inquiry
- Physical science
- Life science
- Earth science

Table 5-17. Framework targets and items by content area: Science

Content area	Targeted percent of items	Proposed number of items	Proposed percent of items
Scientific inquiry	25	5	25
Physical science	25	5	25
Life science	25	5	25
Earth science	25	5	25
Total	100	20	100

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

The proposed test form matches the content category specifications.

Item Overlap

Table 5-18 shows the overlap of items with the ECLS-K. As stated in the proposal, 80 percent of the form items in each domain were to be drawn from the ECLS-K, with 20 percent of the items newly developed. Since the ECLS-K general knowledge test used for kindergarten and first grade was a mix of science and social studies items, there were not enough legacy science items suitable for kindergarten to meet the 80 percent target. More new items were required to meet the science specifications. Thus, only 55 percent of the items were drawn from the ECLS-K.[12]

Table 5-18. Number of items in proposed kindergarten assessment overlapping with the ECLS-K: Science

Assessment	Number of unique overlapping items	Percent of unique overlapping items
ECLS-K K-1, ECLS-K 3rd grade, and/or ECLS-B national	11	55

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

5.3.4 Spanish Early Reading Skills

The ability estimate for the ECLS-K:11 Spanish field test, from the pooled analysis above, was used to estimate the range of abilities that can be expected in the ECLS-K:11 national sample at

[12] It is anticipated that the targeted percentage of overlap with the ECLS-K should be achieved in later rounds.

kindergarten The SERS sample showed a mean theta of -1.23 with a standard deviation of 0.84.[13] Assuming the SERS field test sample is reasonably representative of the national, the range from roughly two standard deviations below the mean (-2.91), to two standard deviations above the mean (+0.45), should include about 95 percent of the K11 national sample.

As with the other domains, as a precaution against encountering ceiling effects, some items with b parameters above +0.45 at the high end have been included. Conversely, to avoid a floor effect items with b parameters below -2.91 would be recommended for inclusion as a safety precaution. However, of the items that functioned well for the SERS, the lowest b parameter was -2.86, just about two standard deviations below the mean. It is not anticipated that numerous children with thetas below this level will be observed. Table 5-19 shows the number of items in difficulty ranges incrementally by standard deviation. Items greater than two standard deviations above the mean are a result of the intentional addition of items to extend difficulties beyond the peak range, and the addition of items to provide overlap between forms needed to support development of a common score scale with the English assessment.)

Table 5-19. Number of items in proposed kindergarten assessment by difficulty range, SERS

Difficulty range	Number of items
Entire assessment	31
Less than -2 s.d.	0
-2 s.d. to -1 s.d.	6
-1 s.d. to mean	8
Mean to +1 s.d.	7
+1 s.d. to +2 s.d.	5
Greater than +2 s.d.	5

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

Other Factors Considered in Form Design

As with the designs of the other assessments, the following additional factors were also considered in assembling the test forms.

[13] Note that the Spanish field test calibration is not directly comparable with the English field test calibration as they are computed on different scales. (The Spanish field test IRT calibration excluded the 2nd and 3rd grade rounds of data.)

Framework

The framework design for the SERS follows those developed for reading as described above. The SERS consists entirely of individual basic skills and vocabulary items, rather than including reading passages and their associated items which may have greater potential for differences in item functioning due to translation. Thus, content categories LOC, INT, and CRIT would be excluded from the design of the SERS. Table 5-20 shows the resulting framework targets and items by content area for the SERS. The first column lists the two content categories included in the SERS. The next column shows the content percentages targeted by the framework. Since only the BS and VOC categories are included in the SERS, the targeted percentages for the SERS were adjusted accordingly. Proposed number and percent of items are listed for the SERS items in the last two columns.

Table 5-20. Framework targets and items by content area: SERS

Content area	Targeted percent of items	Proposed number of items	Proposed percent of items
Basic skills	77	24	77
Vocabulary	23	7	23
Total	100	31	100

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

The majority of Items in the SERS assessment are from the Basic Skills category. Table 5-21 lists the different subcategories of Basic Skills items in the proposed forms, and the number and percent of these items. The phonemic awareness items (substitution, segmentation, blending, etc.), as well as the ending sounds items were excluded from the SERS due to concerns about translation and usage. Although not tasked to match the categories exactly, the percents of items from the English assessment were used to guide the selection of items from the SERS. As a result, the content category percentages in the SERS and the English reading assessment are similar.

Table 5-21. Subcategories of basic skills items in proposed pool: SERS

Basic skills subcategory	Number of items in proposed pool	Percent of items in proposed pool
Letter recognition	5	21
Letter sounds	3	13
Beginning sounds	2	8
Sight words	7	29
Print convention	7	29
Total	24	100

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

Item Overlap

The development of a horizontal, cross-language scale with optimal measurement properties throughout the achievement range calls for items that vary in their difficulty, while at the same time having overlapping items shared across administration languages to permit cross-language scaling. In general, at least 20–30 percent of the items should overlap for adequate scaling. Table 5-22 shows the overlap of items across the proposed SERS assessment and the English reading assessment. Based on the rule of thumb listed above, there are adequate numbers of items across language forms to create a horizontal (cross-language) scale.

Table 5-22. Number of items in proposed kindergarten assessment overlapping with the ECLS-K: SERS

Assessment	Number of unique overlapping items	Percent of unique overlapping items
ECLS-K:2011 Reading	30	36%

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

5.4 Assessment Form Simulations and Review

This section presents the performance simulations used to verify the adequacy of the test forms as well as the technique used to calculate cutscores used to route students to the appropriate second-stage form. Also discussed are the expert and sensitivity reviews conducted.

5.4.1 Assessment Forms and Simulations

Supplemental documents present statistics for the proposed kindergarten assessments in reading, mathematics, science, and the SERS. (The EBRS form statistics are embedded in the reading document.) The documents contain spreadsheets summarizing the selection process, including the items selected for the national assessments, those not selected and why, the ability means and standard deviations by grade, charts of IRT discrimination by difficulty parameters, and the order of items in the forms.

National Sample Performance Simulations and Routing Cutscores

Simulations of routing test and second-stage test performance necessary for setting the cut-points for selection of second-stage forms were computed, and cross-tabulation distributions of routing and second-stage form number right scores were evaluated, to select appropriate routing cutscores for each second-stage form. Specifically 10,000 thetas (ability estimates) were randomly generated, drawn from a normal distribution with mean and standard deviation corresponding to the expected national population statistics, and for each theta, the probability of a correct response was computed for each item on the routing, low, middle, and high forms, separately for each subject.

An estimated number-right score is determined by summing the probabilities for the items on the test form. This, however, never results in a score of zero since, for the multiple-choice items, the probability of a correct response is greater than zero. Therefore, the sum of item probabilities is never zero if there is even a single multiple-choice item. To avoid this limitation on the score calculation and because an integer number-right score is desired for the estimation of cutscores and review of floor and ceiling effects, a random number between 0 and 1 was also generated for each item. If the random number generated was less than or equal to the probability, the item was scored correct (=1); the item was scored incorrect (=0) if the random number was greater than the probability. For example, if the probability for an item, estimated from the item parameters and an individual theta, is .9 and the random number generated is .5, the item would be scored correct. This makes sense because if the probability to correctly answer an item is 90 percent, most times the item should be scored correctly. Conversely, if the probability is .1 and the random number generated is .5, the item should be scored incorrect. Again, since the probability is only 10 percent that this item would be answered correctly, most times the item should be scored incorrectly. Summing the zeros and ones from these calculations resulted in integer scores for each form for each subject. These sums were then cross-tabulated, routing by second-stage form.[14]

Mathematics

Simulations were run on four samples: 1) fall kindergarten – ECLS-K:2011 field test , 2) fall kindergarten – ECLS-K national, 3) spring kindergarten – ECLS-K:2011 field test estimated, and 4) spring kindergarten – ECLS-K national.

[14] Detailed routing score analysis results are available by request as a supplemental document.

Floor and ceiling counts were reviewed. The rule of thumb used to estimate floor effects is to total the number of simulated test takers who would score fewer than three *correct* on both the routing and low forms. If this number is less than three percent of the sample, there is negligible evidence of a floor effect. Similarly, if the total number of test takers scoring fewer than three *incorrect* on both the routing and high forms is less than three percent, there is negligible evidence of a ceiling effect. There is no significant evidence of a floor or ceiling effect using any sample. In addition, review of the counts of simulated test takers who would have fewer than three *incorrect* on the low form and fewer than three *correct* on the middle form, as well as simulated test takers who would have fewer than three *incorrect* on the middle form and fewer than three *correct* on the high form was performed. This is in some ways the opposite of what was discussed in the paragraph above. Here the possibility of a ceiling effect for the routing plus low and floor effect for the routing plus high forms is examined. The routing, low, middle, and high forms were designed to have many items of similar difficulty level to provide for the event that a test taker is routed to a form not appropriate to his or her ability level. These overlapping items provide ample coverage of the ability levels being measured, and allow for the three-form second-stage design.

The approach used to select the optimal cutscores minimized the number of test takers near the edges of score ranges as well as tried to match the number near the lower edge of the routing plus middle score range and the upper routing plus low range for the first cutscore, and the number near the lower edge of the routing plus high score range and the upper routing plus middle range for the second cutscore. The optimal cutscores for mathematics are shown in table 5-23. The score range for the low form was predominantly driven by the fall kindergarten simulations, while the score range for the high form was predominantly driven by the spring simulations, with the middle score range derived from both.

Table 5-23. Cutscores for the ECLS-K:2011 kindergarten assessment in mathematics

Assessment	Routing score that directs to second-stage *low* form	Routing score that directs to second-stage *middle* form	Routing score that directs to second-stage *high* form
Mathematics	0–5	6–13	14–18

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

Reading

The analysis to determine the reading cutscores was the same as mathematics, except that two routing forms were analyzed: 1) router 1, or the set of EBRS items, and 2) router 2, or the set of

additional router items used to differentiate between administration of the middle or high form. Thus, eight simulations were performed, one for each sample by two routing forms. Like in mathematics, there are negligible floor or ceiling effects anticipated for these forms in fall and spring kindergarten. The overall cutscores are listed in tables 5-24 and 5-25.

All children will be administered router 1. For those children proceeding with the remainder of the assessment in English, based on the score ranges in table 5-24, the child will be routed directly to the low second-stage form with a router 1 score of 0–9, or to the additional routing items in router 2 for scores of 10 and above. Those children administered router 2 with a number correct of 0–11 will proceed with the middle form, while those with scores of 12 and above will proceed to the high form.

Table 5-24. Cutscores for the ECLS-K:2011 kindergarten assessment in reading: Router 1

Router	Router 1 score that directs to second-stage *low* form	Router 1 score that directs to router 2
Router 1	0–9	10–20

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

Table 5-25. Cutscores for the ECLS-K:2011 kindergarten assessment in Reading: Router 2

Router	Router 2 score that directs to the second-stage *middle* form	Router 2 score that directs to the second-stage *high* form
Router 2	0–11	12–20

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) field test, fall 2009.

5.4.2 Expert Review of the Assessments

The proposed items for each cognitive domain were reviewed by content experts in each area prior to the field test. The content experts received copies of the proposed items and guidelines for their review. The content reviewer guidelines included the study background, the prior development and testing of the field test items, the objectives of the national tests, and a description of the adaptive two-stage assessment. The guidelines also included specific objectives for each cognitive domain, as well as the following general objectives:

- It is important that the items be accurate: correct content, accurate presentation (language, illustrations, and charts), spelling, and grammar.
- Multiple-choice questions should have a single unambiguously right answer, and all others should be unambiguously wrong.
- Are the open-ended questions fairly easy to evaluate?
- Incorrect options should be plausible responses to the question (i.e., options that could plausibly be selected by a test taker who does not know the answer). Ideally, a test taker who knows the material should get it correct, and one who does not should only be able to guess at random and not be able to eliminate answers that are obviously impossible.
- There should be nothing about the phrasing or the context that is tricky or confusing (e.g., use of metric units in a question that is not trying to measure familiarity with the metric system may be problematic for some students and interfere with their being able to answer the question).
- The question and response formats should not give hints.
- Is the content of the items appropriate and important?
- Are any of the items more characteristic of the typical curriculum of a different grade level?
- Are there ways the presentation (context, language, illustrations, response options) need to be improved?

In addition to these general objectives, there were specific issues to consider for each cognitive domain.

5.4.3 Sensitivity Review

The final items underwent a "sensitivity review" at ETS by a reviewer trained to detect objectionable material such as gender or ethnic stereotyping, inappropriate assumptions about people with disabilities, imbalance of male/female, Black/White, etc., characters in stories or test items, or any other offensive or inappropriate content. No new recommendations resulted from the sensitivity review that had not already been incorporated in the items based on reviews of earlier versions.

6. SCIENCE ACADEMIC RATING SCALE

The fall 2009 English field test tested the feasibility of including a Science Academic Rating Scale (ARS) that could be collected before grade 3. The new Science ARS scores would complement the information collected through the direct child assessment of science knowledge. This chapter describes the performance of the field-tested Science ARS and our recommendations for the national data collection.

We reviewed the performance of the Science ARS in the field test separately by grade. First, we examined the distribution of scores overall and by ability level. We also examined its internal-consistency reliability by grade and means scores by ability level. We then compared the performance of the Science ARS with that of the ARS instruments previously fielded in ECLS-K.

6.1 Kindergarten

Of the 426 Science ARS forms completed by kindergarten teachers, valid achievement level data were obtained for 423. Eighty-five forms (20.1 percent) were for children who were the teachers' highest achieving student; 88 forms (20.8 percent) were for children who were the teachers' lowest achieving student; and 250 (59.1 percent) were for children with average achievement. These results indicate that teachers did not experience difficulties completing the form and that they followed the overall instructions regarding student ability levels.

Means and variances of kindergarten Science ARS item data show good variation, with no floor or ceiling effects. For each item, the full range of possible item values (i.e., 1–5) were used by kindergarten teachers. Table 6-1 shows the average item scores across the completed kindergarten forms, as well as the percentage with the highest possible rating (5), the percentage with the lowest possible rating (1), and the percentage where "not applicable or skill not yet taught" was selected for the item.

Table 6-1. Average item scores across the completed kindergarten Science ARS forms and percent where the highest possible rating, the lowest possible rating, and not applicable or skill not yet taught was selected

Item	Average item score	Percent highest possible rating (5)	Percent lowest possible rating (1)	Percent not applicable or skill not yet taught
K_1. **Uses his/her senses to explore and observe –** for example, observes and notes the habits of classroom pets; identifies environmental sounds; or describes the differences in clay before and after water is added	3.5	23.7	2.6	5.9
K_2. **Forms explanations based on observations and explorations –** for example, describes or draws the conditions (water, soil, sun) that help a plant grow; or explains that a block will slide more quickly down a steeper slope	3.2	15.8	8.7	10.4
K_3. **Classifies and compares living and non-living things in different ways –** for example, classifies objects according to "things that are alive and not alive," or "things that fly and things that crawl," or "plants and animals"	3.6	23.4	8.7	13.2
K_4. **Makes logical predictions when pursuing scientific investigations –** for example, observes and identifies patterns in nature and predicts what happens next (e.g., if told the sky became dark and cloudy, predicts that it will rain; or predicts if a new object will float or sink)	3.3	18.0	3.8	5.2
K_5. **Communicates scientific information –** for example, records or describes the properties of common objects verbally or through drawings or graphs	3.1	14.0	10.7	7.8
K_6. **Demonstrates understanding of physical science concepts –** for example, makes observations that different materials have different properties and that objects are made of different types of materials; compares the relative sizes and characteristics of objects; or describes and explains the different way things move	3.2	17.0	7.6	11.8
K_7. Demonstrates understanding of life science concepts – for example, recognizes the five senses and the related body parts; identifies major structures and functions of parts of plants and animals; or describes the similarities and differences in the appearance and behavior of plants and animals	3.3	17.4	5.4	6.1
K_8. Demonstrates understanding of earth and space science concepts – for example, identifies that changes in weather occur from day to day and season to season; describes properties of rocks, soil, and water; or identifies that the sun gives light and heat to Earth	3.2	16.5	7.3	6.1

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

Mean item scores were approximately in the middle of the range, although K_1 (3.5) and K_2 (3.6) were slightly higher. The percentage of children who were rated with the highest possible score ranged from 14.0 to 23.7 percent. This is higher than the percentage of children who were rated "proficient" in the ECLS-K grade 3 Science ARS scores, which ranged from 8.5 to 12.7 percent across those items. Although items K_1 (23.7 percent) and K_2 (23.4 percent) had relatively higher percentages of children who were scored as proficient, they still showed strong correlations with the total score (*r*s = .84 and .87, respectively). These correlation coefficients are similar in magnitude to those found with the other Science ARS items (*r*s of other items ranged from .84 to .88).

The percentage of children who were rated with the lowest possible score ranged from 2.6 to 10.7. In comparison, the percentage of children who were rated "not yet" in the ECLS-K grade 3 Science ARS scores, which ranged from 2.7 to 5.4 percent across those items.

Internal-consistency reliability was very good for the kindergarten Science ARS items. Cronbach's coefficient alpha was calculated to examine internal-consistency reliability. The alpha coefficient was $\alpha = 0.96$. "Not applicable" was recoded to missing in these reliability analyses. This coefficient is similar in magnitude to the alpha coefficients obtained for the grade 3 Science ARS ($\alpha = .95$) and the various kindergarten ARS (α all above 0.90). Deletion of item K_1 or K_2 does not change the magnitude of the alpha coefficient.

Average Science ARS scale scores significantly increased with achievement levels of rated kindergarten children. Scale scores were calculated by taking the arithmetic mean of the eight items ("not applicable" was recoded to missing prior to calculating scale scores). A one-way analysis of variance (ANOVA) found that scores for rated children differed by performance level (F (2, 420) = 206.9; $p < .0001$). The highest achieving students had the highest scale scores (4.4). The lowest achieving students had the lowest scale scores (2.1). Children with average achievement had scores in the middle (3.3). Post-hoc Scheffé tests found that all three scores significantly differed from each other ($p < .05$).

6.2 First Grade

Of the 423 Science ARS forms completed by first-grade teachers, valid achievement level data were obtained for 415 forms. Eighty-three forms (20.0 percent) were for children who were the

teachers' highest achieving student; 91 forms (21.9 percent) were for children who were the teachers' lowest achieving student; and 241 (58.1 percent) were for children with average achievement.

Means and variances of first-grade Science ARS item data showed good variation. For each item, the full range of possible item values (i.e., 1–5) were used by first-grade teachers. As shown in table 6-2, mean item scores were approximately in the middle of the range. The percentages of children who were rated with the highest possible score ranged from 6.7 to 19.7. This range overlaps the range of percentages of children who were rated "proficient" across the ECLS-K grade 3 Science ARS items which ranged from 8.5 to 12.7 percent. Although F_1 had a relatively higher percentage of children who were rated as proficient, it still showed strong correlations with the total score (r = .80). It should be noted that this correlation coefficient is smaller in magnitude to those found with the other Science ARS items (rs of other items ranged from .84 to .88).

The percentages of children who were rated with the lowest possible score ranged from 3.6 to 10.0. In comparison, the percentage of children who were rated "not yet" in the ECLS-K grade 3 Science ARS scores ranged from 2.7 to 5.4 percent across those items. Two items had relatively higher percentages of children who had not yet had the respective skill taught at the time of the rating: F_6 (46.1 percent) and F_8 (33.3 percent).

Table 6-2. Average item scores across the completed first-grade Science ARS forms and percent where the highest possible rating, the lowest possible rating, and not applicable or skill not yet taught was selected

Item		Average item score	Percent highest possible rating (5)	Percent lowest possible rating (1)	Percent not applicable or skill not yet taught
F_1.	**Uses his/her senses to explore and observe –** for example, moves objects and describes how a push or pull can change the way an object is moving; or observes that some living things closely resemble their parents; or observes and describes properties of rocks, soil, and water; or uses tools (such as hand lenses, thermometers, rulers) to gather information about objects around them	3.4	19.7	3.6	2.6
F_2.	**Forms explanations based on observations and explorations –** for example, explains the best growing conditions for a plant after investigating with light and water; or concludes that earthworms come out of the soil because it's raining after paying attention to the sidewalks on a rainy day	3.1	12.6	8.1	7.6
F_3.	**Classifies and compares living and non-living things in different ways –** for example, classifies vegetables that grow above or below the ground; classifies different sounds as either low pitch or high pitch; or measures objects and classifies them by size or weight	3.3	14.5	6.2	14.7
F_4.	**Makes logical predictions when pursuing scientific investigations –** for example, predicts whether or not objects are magnetic based on the materials they are made of	2.9	8.1	10.0	20.4
F_5.	**Communicates scientific information –** for example, records data from measurement tools (e.g., clocks, thermometers, etc.) or constructs bar graphs	3.0	11.4	7.1	5.9
F_6.	**Demonstrates understanding of physical science concepts –** for example, identifies the three states of matter; identifies that heat causes change and compares objects according to temperature; or compares the way different objects move (in straight line, by vibration, in a circle)	3.0	6.7	8.3	46.1
F_7.	**Demonstrates understanding of life science concepts –** for example, understands that living organisms inhabit various environments and have various external features to help them satisfy their needs; differentiates between those living things that closely resemble their parents (e.g., chick) and those living things that do not (e.g., tadpole); or recognizes that all plants and animals have basic life needs (e.g., air, water, food, etc.)	3.4	15.4	4.3	13.5
F_8.	**Demonstrates understanding of earth and space science concepts –** for example, describes how weather affects people's daily activities, describes how land and water store heat from the sun and then warm the air over the land and water; explains that shadows are caused when sunlight is blocked by objects; or identifies natural resources	3.1	10.5	8.1	33.3

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011)English field test, fall 2009.

Internal-consistency reliability was very good for the first-grade Science ARS items. An alpha coefficient of $\alpha = 0.96$ indicated that there was very good internal-consistency reliability in the first-grade data, similar to those found in ARS data from the ECLS-K. Deletion of item F_1 does not change the magnitude of the alpha coefficient.

Average Science ARS scale scores significantly increased with achievement levels of rated first-grade children. Similar to the analyses with kindergarten children, first-grade scale scores were calculated by taking the arithmetic mean of the eight items ("not applicable" was recoded to missing prior to calculating scale scores). A one-way ANOVA found that scores for rated first-grade children differed by performance level (F (2, 414) = 166.5; $p < .0001$). The highest achieving students had the highest scale scores (4.1). The lowest achieving students had the lowest scale scores (2.0). Children with average achievement had scores in the middle (3.2). Post-hoc Scheffé tests found that all three scores significantly differed from each other ($p < .05$).

6.3 Second Grade

Of the 390 Science ARS forms completed by second-grade teachers, valid achievement level data were obtained for 384 forms. Eighty forms (20.8 percent) were for children who were the teachers' highest achieving student; 80 forms (20.8 percent) were for children who were the teachers' lowest achieving student; and 224 (58.3 percent) were for children with average achievement.

Means and variances of second-grade Science ARS item data showed good variation. For each item, the full range of possible item values (i.e., 1–5) were used by second-grade teachers. Mean item scores were approximately in the middle of the range (table 6-3). The percentage of children who were rated with the highest possible score ranged from 3.9 to 26.1 percent. S_1 had 26.1 percent of children as proficient but still showed strong correlations with the total score ($r = .70$), although this correlation coefficient is smaller in magnitude to those found with the other Science ARS items (rs of other items ranged from .83 to .88).

The percentages of children who were rated with the lowest possible score, ranged from 2.6 to 9.3. In comparison, the percentage of children who were rated "not yet" in the ECLS-K grade 3 Science ARS scores ranged from 2.7 to 5.4 percent across those items.

Two items had relatively higher percentages of second-grade children who had not yet had the respective skill taught at the time of the rating: S_6 (47.8 percent) and S_8 (33.1 percent).

Table 6-3. Average item scores across the completed second-grade Science ARS forms and percent where the highest possible rating, the lowest possible rating, and not applicable or skill not yet taught was selected

Item	Average item score	Percent highest possible rating (5)	Percent lowest possible rating (1)	Percent not applicable or skill not yet taught
S_1. **Uses his/her senses to explore and observe –** for example, compares and classifies objects according to two or more physical attributes (e.g., a basketball is round and has a rough texture, a feather is soft and is 7 centimeters long); or uses observations through the senses to predict an outcome of a simple investigation such as a marble will roll with a greater speed if a ramp is raised 2 cm	3.6	26.1	2.6	1.8
S_2. **Forms explanations based on observations and explorations –** for example, explains why one boat floats and another does not; or concludes that a candle stays lit longer under a larger jar because there is more oxygen available; or explains how many layers of clothing provide insulation against heat loss	3.1	12.9	7.2	6.2
S_3. **Classifies and compares living and non-living things in different ways –** for example, compares living things based on life cycle; or compares mixtures based on size and/or substance; or describes differences in how the environment affects living things (e.g., migration of birds as the availability of food becomes less when autumn changes to winter) versus how it affects non-living things (e.g., erosion of rocks, evaporation of water)	3.3	12.9	3.9	10.3
S_4. **Makes logical predictions when pursuing scientific investigations –** for example, predicts the outcome of a simple investigation and compares result with prediction, such as predicting if a plant will grow best in direct sunlight or in shade	3.3	16.7	4.6	4.4
S_5. **Communicates scientific information –** for example, records data gathered using simple equipment in simple investigations (e.g., changes in weather conditions); summarizes data using charts or graphs; or uses correct units of measurement when recording or summarizing data	3.2	12.9	6.2	10.3
S_6. **Demonstrates understanding of physical science concepts –** for example, describes the effects of electrically charged materials and magnets; or explains that sound is made by vibrating objects and describe its pitch and loudness	2.7	3.9	8.2	47.8

See note at end of table.

Table 6-3. Average item scores across the completed second-grade Science ARS forms and percent where the highest possible rating, the lowest possible rating, and not applicable or skill not yet taught was selected—Continued

Item	Average item score	Percent highest possible rating (5)	Percent lowest possible rating (1)	Percent not applicable or skill not yet taught
S_7. **Demonstrates understanding of life science concepts –** for example, explains that the sequential stages of life cycles are different for different animals, describes how living organisms depend on each other and their environments for survival; identifies differences between living and nonliving objects; or describes how the environment influences some characteristics of living organisms.....	3.2	12.6	4.6	22.9
S_8. **Demonstrates understanding of earth and space science concepts –** for example, describes the effects of weathering and erosion; the relationship between the Sun and the Earth; the use of tools to measure weather conditions; or the processes involved with soil formation ...	2.9	8.3	9.3	33.1

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) English field test, fall 2009.

Internal-consistency reliability was very good for the second-grade Science ARS items. Internal-consistency reliability of the second-grade Science ARS data was very good with an alpha coefficient $\alpha = 0.95$, similar to those in the ECLS-K ARS data. Deletion of item S_1 does not change the magnitude of the alpha coefficient.

Average Science ARS scale scores significantly increased with achievement levels of rated second-grade children. A one-way ANOVA found that average scores for rated second-grade children differed by performance level (F (2, 382) = 194.8; $p < .0001$). The highest achieving students had the highest scale scores (4.2). The lowest achieving students had the lowest scale scores (2.0). Children with average achievement had scores in the middle (3.3). Post-hoc Scheffé tests found that all three scores significantly differed from each other ($p < .05$).

6.4 Summary and Recommendations

Generally, average item scores were in the middle of the possible range and showed good variation with teachers using the entire range of scores. Some items, most notably K_1, F_1, and S_1 had higher percentages of children rated as proficient, although they showed strong correlations with the total

score. Their deletions from their respective scales do not impact the internal consistency of the scale. Two items in first grade and second grade tapped science skills or knowledge (demonstrates understanding of physical science concepts and demonstrates understanding of earth and space science concepts) that had higher frequencies of not being taught yet to the rated student, which may not be surprising in the beginning of the school year. Data across all three grades showed very strong internal consistency reliability, as well as expected differences across achievement levels of students.

Based on these findings, the kindergarten, first-grade, and second-grade Science ARS instruments are recommended for fielding in the ECLS-K:2011. In their review of the instruments (prior to this report) NCES has requested that the number of examples listed under each item be shortened to two examples to ease the burden on respondents, which were made before fielding these scales.

APPENDIX B:
IRT ESTIMATION USING PARSCALE

This appendix provides more detail on how the raw item responses are prepared for use in PARSCALE, how PARSCALE estimates the IRT model parameters, and what quality control checks are performed on the PARSCALE estimation output.

Preparing Data Files for PARSCALE

The first step in processing children's raw item responses was preparing scored-item files for use in the IRT calibration procedures. These files were first prepared separately for each round of data collection, fall and spring kindergarten. 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 either refused by the child or multiple choice items with responses of don't know 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 some instances, discontinue rules were employed such that the more difficult items at the end of the assessment were not administered if a child had performed poorly on the easier items earlier on. The "not reached" or "not administered" code was used for items that were not answered by an individual child for any of the following reasons:

- The item was presented on a test form that the child was not administered (e.g., the child was routed to the middle second-stage form and the item appeared only on the high form).
- The item appeared on the form subsequent to the enforcement of a discontinue rule.
- The child was unable to complete the assessment and the item was not reached.

The quality control procedure for confirming that the processing of the prepared data files 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 possible conversions found in the raw data file could be checked. For example, if the spaced sample of quality control records happened to have only data for children who were routed

to the low and middle second-stage forms, additional records were reviewed so that score conversions for children routed to the high second-stage form 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. 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 round of collection separately, the scored-item files (from both rounds, fall and spring) were stacked, and frequency counts were checked to confirm the accuracy of the concatenated files. The non–IRT-based scores were computed at this time and then visually checked for accuracy in the same spaced sample. These number-right scores were included in the scored-item files for additional quality control purposes.

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 rounds), the frequency counts were checked against the aggregates of the frequencies for the separate forms in which the items originally appeared. These frequency counts, and item means computed on the verified scored-item files, provided the basis for checking the results of the IRT scaling steps.

PARSCALE Estimate of the IRT Model

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 and item priors.[1] 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. When the PARSCALE program is applied to dichotomously scored items, 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 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:

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

$$L(\beta) = \prod_g \prod_{j:g} \int_\theta P(x_{j:g} \mid \theta, \beta)\, f_g(\theta) d(\theta)$$
$$\approx \prod_g \prod_{j:g} \Sigma_k P(x_{j:g} \mid \theta = X_k, \beta)\, A_g(X_k). \quad (1)$$

In equation (1), $L(\beta)$ 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 θ 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 latter is used in the ECLS-K:2011 calibrations.)

The $f_g(\theta)$ in Equation 1 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 1 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., $A_g(X_k) = f_g(X_k) / \Sigma_k f_g(X_k)$). In the ECLS-K:2011, each round of data collection (i.e., fall and spring kindergarten) is treated as a separate population for calibration; thus, the more general population density function is used.

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 (expectation) 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 (maximization) 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 constrain the priors to be normal or allow their shape to vary. The ECLS-K:2011 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 1 finds those item parameters that maximize the likelihood across two points (fall and spring). 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,[2] 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.

Quality Control for PARSCALE Estimation

Statistics and graphs produced by the PARSCALE program and an IRT 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

[2] 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 standard error of measurement *(SEM)* are reported for all eligible children on the data file.

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 estimates demonstrated the cohesiveness of the *whole set* of items used in each domain across the assessments. High *a* parameter estimates (1.0 or above) were found for items strongly related to the underlying construct represented by the item pool.

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 were superimposed on the same graph. The closeness of fit of the logistic function to the data can be interpreted as confirming the appropriateness of the IRT model for scoring the tests.

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 round. These scores were checked by printing a spaced sample of every 1,000th data 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 expectations. For the scale scores, means were expected to increase from round to round, 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, that child's predicted IRT scale score had the potential to indicate correct responses for all items).

This page intentionally left blank.

APPENDIX C

ECLS-K:2011 KINDERGARTEN IRT ITEM PARAMETERS

Table C-1. ECLS-K:2011 Kindergarten reading IRT item parameters: School year 2010–11

Item	Test form(s)	a[1]	b[2]	c[3]
Vocabulary #1	EBRS	1.17704	-2.91447	0.00001
Vocabulary #2	EBRS	1.37321	-2.40472	0.00001
Letter Recognition #1	EBRS	1.70408	-2.15674	0.13627
Letter Recognition #2	EBRS	1.71257	-1.5952	0.00001
Letter Recognition #3	EBRS	1.96998	-1.58179	0.00001
Letter Recognition #4	EBRS	1.88319	-1.57335	0.00001
Letter Recognition #5	EBRS	1.6943	-1.64979	0.00001
Letter Sounds #1	EBRS	1.38581	-2.07573	0.00001
Letter Sounds #2	EBRS	1.55072	-1.46718	0.00001
Phonemic Awareness #1	EBRS	0.53143	-1.78423	0.10716
Phonemic Awareness #2	EBRS	0.52106	-1.0201	0.03114
Beginning Sounds #1	EBRS	0.70565	-1.16555	0.00001
Beginning Sounds #2	EBRS	1.22949	-0.96633	0.00001
Beginning Sounds #3	EBRS	1.13817	-0.95741	0.00001
Beginning Sounds #4	EBRS	0.73804	-0.36825	0.00001
Beginning Sounds #5	EBRS	0.74939	0.02104	0.00001
Ending Sounds #1	EBRS	0.96905	-0.48041	0.00001
Ending Sounds #2	EBRS	0.7656	0.17274	0.00001
Ending Sounds #3	EBRS	0.88427	0.13375	0.00001
Sight Words #1	EBRS	1.71919	0.14585	0.00001
Blending #1	Router 2	1.42907	-0.02898	0.00001
Blending #2	Router 2	1.25056	0.24022	0.00001
Phonemic Awareness #3	Router 2	1.00267	0.32451	0.00001
Phonemic Awareness #4	Router 2	1.55809	0.29642	0.00001
Phonemic Awareness #5	Router 2	0.92075	0.53036	0.00001
Phonemic Awareness #6	Router 2	0.78028	1.22763	0.00001
Phonemic Substitution #1	Router 2	1.07277	0.17904	0.00001
Phonemic Substitution #2	Router 2	1.09014	0.71764	0.00001
Segmentation #1	Router 2	1.22215	-0.4724	0.00001
Segmentation #2	Router 2	0.60328	1.31526	0.00001
Rhyming #1	Router 2	0.80675	0.6519	0.00001
Rhyming #2	Router 2	0.76021	1.00692	0.00001

See notes at end of table.

Table C-1. ECLS-K:2011 Kindergarten reading IRT item parameters: School year 2010–11—Continued

Item	Test form(s)	a[1]	b[2]	c[3]
Sight Words #2	Router 2	1.88653	0.53377	0.00001
Sight Words #3	Router 2	1.77918	0.60113	0.00001
Sight Words #4	Router 2	1.84442	0.87492	0.00001
Sight Words #5	Router 2	1.69069	0.92875	0.00001
Locate/Recall #1	Router 2	2.66323	0.93631	0.1841
Locate/Recall #2	Router 2	1.53576	0.86918	0.2411
Locate/Recall #3	Router 2	2.70981	0.98097	0.19268
Locate/Recall #4	Router 2	2.31987	1.26422	0.22963
Print Convention #1	Low	1.42433	-2.42719	0.00001
Print Convention #2	Low	1.58715	-2.20039	0.00001
Print Convention #3	Low	1.27247	-1.51626	0.00001
Print Convention #4	Low	1.25022	-1.37244	0.00001
Letter Recognition #6	Low	1.2937	-1.79252	0.14113
Print Convention #5	Low, Middle	1.14589	-1.86629	0.00001
Letter Sounds #3	Low, Middle	0.95171	-1.57989	0.00361
Letter Sounds #4	Low, Middle	0.70478	-1.61197	0.02559
Ending Sounds #4	Low, Middle, High	0.82608	-0.45337	0.00001
Vocabulary #3	Low	1.11858	-2.42266	0.2901
Vocabulary #4	Low	1.19177	-1.88342	0.09453
Vocabulary #5	Low	0.8775	-1.33308	0.12518
Vocabulary #6	Low, Middle	0.71379	-1.15959	0.06834
Vocabulary #7	Low	0.89072	-0.96438	0.16193
Print Convention #6	Low, Middle	0.69349	-0.93563	0.00001
Print Convention #7	Low, Middle	0.81458	-0.48552	0.00001
Print Convention #8	Low, Middle	0.84407	0.1791	0.00001
Vocabulary #8	Middle	0.41873	-0.45989	0.1424
Vocabulary #9	Middle	0.47952	0.89666	0.22515
Vocabulary #10	Middle	0.46668	1.24699	0.25344
Sight Words #6	Middle	1.52641	-0.05398	0.00001
Sight Words #7	Middle	1.98265	0.46795	0.00001
Sight Words #8	Middle	1.5745	1.10574	0.00001
Locate/Recall #5	High	3.49257	1.34734	0.2379
Locate/Recall #6	High	5.39753	1.45029	0.20238
Locate/Recall #7	High	4.95035	1.53376	0.23953
Locate/Recall #8	High	4.49346	1.58458	0.1752
Sight Words #9	High	4.33504	1.19747	0.00001

See notes at end of table.

Table C-1. ECLS-K:2011 Kindergarten reading IRT item parameters: School year 2010–11—Continued

Item	Test form(s)	a[1]	b[2]	c[3]
Sight Words #10	High	5.01628	1.23992	0.00001
Sight Words #11	High	3.46833	1.54382	0.00001
Sight Words #12	High	2.73852	1.8724	0.00001
Locate/Recall #9	High	1.87727	1.08348	0.51166
Locate/Recall #10	High	3.58965	1.23897	0.35521
Locate/Recall #11	High	5.44207	1.43971	0.00001
Integrate/Interpret #1	High	3.19862	1.29249	0.31798
Critique/Evaluate #1	High	2.04732	1.81853	0.25157
Integrate/Interpret #2	High	3.11846	1.79089	0.00001
Locate/Recall #12	High	3.92307	1.90426	0.00001
Locate/Recall #13	High	3.06075	1.75225	0.00001
Locate/Recall #14	High	3.05144	1.7149	0.00001
Integrate/Interpret #3	High	2.72899	1.83576	0.00001
Vocabulary #11	High	2.16822	2.35933	0.00001
Critique/Evaluate #2	High	1.6831	2.41931	0.00001

[1] Item Response Theory (IRT) discrimination parameter.
[2] Item Response Theory (IRT) difficulty parameter.
[3] Item Response Theory (IRT) guessing parameter.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2010 and spring 2011.

Table C-2. ECLS-K:2011 kindergarten mathematics IRT item parameters: School year 2010–11

Item	Test form(s)	a[1]	b[2]	c[3]
Data Analysis #1	Routing	0.58826	-1.63275	0.00717
Geometry #1	Routing	0.66590	-2.36571	0.60000
Measurement #1	Routing	0.82477	-1.03543	0.28113
Number Sense - Addition/Subtraction #1	Routing	0.90777	0.03027	0.00732
Number Sense - Counting #1	Routing	0.94957	-0.89364	0.00713
Patterns #1	Routing	0.96218	0.74949	0.26163
Number Sense - Addition/Subtraction #2	Routing	0.98742	0.30492	0.00729
Number Sense - Addition/Subtraction #3	Routing	0.99228	0.31862	0.00766
Number Sense - Addition/Subtraction #4	Routing	1.02309	0.98404	0.00709
Number Sense - Addition/Subtraction #5	Routing	1.09743	0.38372	0.00725
Number Sense - Addition/Subtraction #6	Routing	1.24458	0.67093	0.17163
Patterns #2	Routing	1.26978	1.00727	0.00717
Number Sense - Ordinality	Routing	1.27000	-0.16529	0.00730
Number Sense - Number Recognition #1	Routing	1.29456	-1.17058	0.00718
Number Sense - Sequencing #1	Routing	1.36852	-0.43561	0.00726
Number Sense - Number Recognition #2	Routing	1.41898	-0.12230	0.00718
Number Sense - Addition/Subtraction #7	Routing	1.45572	0.89732	0.00751
Number Sense - Number Recognition #3	Routing	1.51363	-2.27465	0.00718
Number Sense - Fewer/More #1	Low	0.32665	-0.66772	0.00718
Number Sense - Counting #2	Low	0.45374	-4.69026	0.00717
Number Sense - Counting #3	Low	0.46608	-2.44108	0.25882
Number Sense - Fewer/More #2	Low	0.47875	0.09545	0.00718
Number Sense - Fewer/More #3	Low	0.47977	-1.04764	0.46644
Number Sense - Counting #4	Low	0.73313	-0.87309	0.00717
Data Analysis #2	Low	0.79346	-2.42697	0.00718
Number Sense - Counting #5	Low	0.81280	-2.66617	0.00717
Number Sense - Fewer/More #4	Low	0.86728	-1.14506	0.00718
Number Sense - Counting #6	Low	1.21022	-0.84371	0.15579
Number Sense - Number Recognition #4	Low	1.32963	-1.78673	0.00718
Number Sense - Counting #7	Low	1.39015	-2.17168	0.00718
Number Sense - Addition/Subtraction #8	Low, Middle	0.71490	-1.22448	0.00720
Data Analysis #3	Low, Middle	0.80133	-1.55796	0.00716
Data Analysis #4	Low, Middle	0.80706	-1.92272	0.00716
Measurement #2	Low, Middle	0.88961	-1.12755	0.27931

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten mathematics IRT item parameters: School year 2010–11—Continued

Item	Test form(s)	a[1]	b[2]	c[3]
Number Sense - Number Recognition #5	Low, Middle, High	1.32987	-0.47767	0.00714
Number Sense - Counting #8	Middle	0.60259	-0.65709	0.13093
Number Sense - Fewer/More #5	Middle	0.64178	-0.19626	0.21673
Geometry #2	Middle	0.65623	-0.13244	0.17536
Number Sense - Fewer/More #6	Middle	0.67810	0.21172	0.12250
Number Sense - Addition/Subtraction #9	Middle	0.68746	-0.18220	0.08315
Patterns #3	Middle	0.78540	-0.21601	0.27445
Number Sense - Estimation #1	Middle	0.87449	0.12913	0.15758
Number Sense - Number Recognition #6	Middle	0.99815	-0.36646	0.20174
Number Sense - Addition/Subtraction #10	Middle	1.04225	0.38229	0.00717
Number Sense - Addition/Subtraction #11	Middle	1.07366	0.09830	0.12034
Number Sense - Counting #9	Middle	1.46858	0.82267	0.00712
Number Sense - Sequencing #2	Middle	1.54550	0.72402	0.00717
Number Sense - Addition/Subtraction #12	Middle	3.73736	0.21493	0.00728
Patterns #4	Middle, High	1.03496	0.54848	0.29342
Number Sense - Addition/Subtraction #13	Middle, High	1.08845	1.61394	0.01570
Number Sense - Addition/Subtraction #14	Middle, High	2.06009	0.44219	0.00703
Number Sense - Addition/Subtraction #15	Middle, High	2.44748	1.65596	0.00669
Number Sense - Addition/Subtraction #16	Middle, High	2.64160	0.13706	0.00736
Patterns #5	High	0.69102	-0.10694	0.19717
Number Sense - Addition/Subtraction #17	High	0.74424	-0.43912	0.00717
Data Analysis #5	High	0.80023	2.35816	0.00718
Number Sense - Multiplication/Division #1	High	1.09912	2.00999	0.00718
Patterns #6	High	1.19670	1.34231	0.00716
Number Sense - Addition/Subtraction #18	High	1.28066	0.44528	0.00717
Number Sense - Money #1	High	1.42507	2.39731	0.00728
Number Sense - Addition/Subtraction #19	High	1.43360	1.72465	0.00717
Number Sense - Estimation #2	High	1.55580	2.06217	0.00723
Data Analysis #6	High	1.59600	1.57863	0.00720
Number Sense - Multiplication/Division #2	High	1.64690	1.48735	0.00720
Patterns #7	High	1.74515	0.83431	0.00717

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten mathematics IRT item parameters: School year 2010–11—Continued

Item	Test form(s)	a[1]	b[2]	c[3]
Number Sense - Addition/Subtraction #20	High	1.76243	1.12956	0.00715
Number Sense - Addition/Subtraction #21	High	1.81985	1.83200	0.00721
Number Sense - Addition/Subtraction #22	High	1.89653	0.87868	0.00716
Number Sense - Addition/Subtraction #23	High	1.90998	1.76027	0.00713
Number Sense - Addition/Subtraction #24	High	1.91524	1.74505	0.00712
Number Sense - Multiplication/Division #3	High	2.00911	1.75329	0.00726
Patterns #8	High	2.10658	0.92751	0.00717
Number Sense - Multiplication/Division #4	High	2.26105	1.68549	0.00725
Number Sense - Addition/Subtraction #25	High	2.66354	1.96964	0.00711
Number Sense - Money #2	High	2.81439	2.19963	0.00708

[1] Item Response Theory (IRT) discrimination parameter.
[2] Item Response Theory (IRT) difficulty parameter.
[3] Item Response Theory (IRT) guessing parameter.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2010 and spring 2011.

Table C-3. ECLS-K:2011 kindergarten science IRT item parameters: School year 2010–11

Item	a[1]	b[2]	c[3]
Life Science #1	1.22028	-1.12805	0.825
Physical Science #1	0.59425	-0.25402	0.56292
Earth Science #1	0.55531	-1.56503	0.40063
Scientific Inquiry #1	0.73863	-1.16655	0.00001
Life Science #2	1.17529	-1.09259	0.00001
Earth Science #2	0.93555	-0.69703	0.00001
Earth Science #3	0.78569	-0.84103	0.00001
Physical Science #2	1.06078	-0.09729	0.16824
Physical Science #3	0.83743	-0.28915	0.19376
Scientific Inquiry #2	0.79794	-0.88006	0.00001
Life Science #3	1.08132	0.40997	0.00001
Scientific Inquiry #3	0.85499	0.43745	0.30835
Life Science #4	0.76793	0.81077	0.39703
Scientific Inquiry #4	0.91231	0.5512	0.14747
Scientific Inquiry #5	0.87386	3.06918	0.21654
Physical Science #4	1.36161	1.27296	0.16679
Physical Science #5	1.07314	1.86303	0.05265
Earth Science #4	1.14704	1.60448	0.13601
Earth Science #5	1.17355	1.59889	0.2286
Life Science #5	0.84546	1.44217	0.28632

[1] Item Response Theory (IRT) discrimination parameter.
[2] Item Response Theory (IRT) difficulty parameter.
[3] Item Response Theory (IRT) guessing parameter.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2010 and spring 2011.

Table C-4. ECLS-K kindergarten Spanish early reading skills IRT item parameters: School year 2010–11

Item	a[1]	b[2]	c[3]
Print Convention #1	0.81797	-2.13719	0.00001
Vocabulary #1	0.78887	-2.20602	0.00001
Vocabulary #2	0.93569	-1.58347	0.00001
Print Convention #2	0.98202	-0.87133	0.00001
Print Convention #3	0.82452	-1.64153	0.00001
Print Convention #4	0.92909	-0.52728	0.00001
Print Convention #5	0.71434	0.2762	0.00001
Print Convention #6	0.7575	0.83247	0.00001
Print Convention #7	0.90908	0.18915	0.00001
Letter Recognition #1	0.3855	-0.14996	0.26249
Letter Recognition #2	1.80561	0.00632	0.00001
Letter Recognition #3	1.96161	-0.21563	0.00001
Letter Recognition #4	2.23397	-0.04214	0.00001
Letter Recognition #5	2.02146	-0.27069	0.00001
Letter Sounds #1	1.15075	-0.8699	0.00001
Letter Sounds #2	0.75017	0.16183	0.23992
Letter Sounds #3	0.86123	0.10971	0.21788
Beginning Sounds #1	1.00379	0.00331	0.00001
Beginning Sounds #2	0.80781	0.11271	0.00001
Vocabulary #3	0.4814	0.05114	0.31858
Vocabulary #4	0.60961	-0.68666	0.26609
Vocabulary #5	0.61845	1.45587	0.27612
Vocabulary #6	0.9742	0.89897	0.19706
Vocabulary #7	0.74188	1.62755	0.2947
Sight Words #1	2.46581	0.70649	0.00001
Sight Words #2	1.85303	0.63806	0.00001
Sight Words #3	5.13953	0.88729	0.00001
Sight Words #4	3.81138	1.18548	0.00001
Sight Words #5	5.88108	0.94424	0.00001
Sight Words #6	5.15731	1.08817	0.00001
Sight Words #7	5.57559	1.14513	0.00001

[1] Item Response Theory (IRT) discrimination parameter.
[2] Item Response Theory (IRT) difficulty parameter.
[3] Item Response Theory (IRT) guessing parameter.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2010 and spring 2011.