NCES 2018-183 U.S. DEPARTMENT OF EDUCATION

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

First-Grade and Second-Grade Psychometric Report

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

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

First- and Second-Grade 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 2018-183
U.S. Department of Education

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-IES-10-C-0048 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), First- and Second-Grade Psychometric Report* (NCES 2018-183). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved from https://nces.ed.gov/pubsearch.

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

CONTENTS

CONTENTS—Continued

Chapter Page

CONTENTS—Continued

CONTENTS—Continued

CONTENTS—Continued

Chapter Page

CONTENTS—Continued

List of Appendixes

CONTENTS—Continued

List of Tables

CONTENTS—Continued

List of Tables—Continued

CONTENTS—Continued

List of Tables—Continued

CONTENTS—Continued

List of Tables—Continued

CONTENTS—Continued

List of Tables—Continued

CONTENTS—Continued

List of Tables—Continued

CONTENTS—Continued

List of Tables—Continued

CONTENTS—Continued

List of Tables—Continued

CONTENTS—Continued

List of Tables—Continued

List of Figures

CONTENTS—Continued

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 first- and second-grade data collections.[1] The focus of this volume is the third through sixth rounds of data collection: the fall 2011 and spring 2012 first-grade and the fall 2012 and spring 2013 second-grade rounds. Readers interested in the kindergarten year should refer to the *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), hereinafter referred to as the base-year user's manual, and the *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) Kindergarten Psychometric Report* (Najarian et al. 2018). 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

[1] Although the study refers to later rounds of data collection by the grade the majority of children are expected to be in (that is, the modal grade for children who were in kindergarten in the 2010–11 school year), children are being included in subsequent data collections regardless of their grade level. Children may not be in the modal grade due to retention in a grade or promotion to a higher grade ahead of schedule.

draws together information from multiple sources, including children, parents, teachers, school administrators, and early care and education providers, 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 First-Grade and Second-Grade Data Collections

The ECLS-K:2011 provides national data on elementary school students' characteristics as they progressed from kindergarten in the 2010–11 school year through the spring of 2016 when most of the students 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 kindergartners enrolled in approximately 970 schools across the United States.[2] During both the first- and second-grade years, as in kindergarten, there were two data collections: one at the beginning (fall) and one near the end (spring) of the school year.

[2] 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.

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.

Only base-year (i.e., kindergarten) respondents[3] were eligible for the first- and second-grade data collections. However, an eligible student did not have to participate in first grade to be eligible for data collection in second grade. The fall data collections in first and second grade were conducted with only a subsample of all students who were base-year respondents. This subsample included approximately 6,110 students attending schools within 30 PSUs that had been sampled from the full group of 90 PSUs selected for the study in the base year. The sample for the spring first-grade and spring second-grade data collections included all 18,174 base-year respondents.

Some students initially included in the first- and second-grade samples were excluded from actual data collection. Students who were initially included in the first- and second-grade samples were eligible for actual data collection if they were living in the United States at the time of data collection. Due to the increased data collection costs associated with following students who transferred from their original sample school (referred to as movers), only a subsample of movers were followed into their new schools. Once a mover was subsampled out of the study, he or she was not contacted to participate in any subsequent data collections. Although information was not collected from all students in every round as described above, the study sampling procedures, combined with application of proper statistical techniques during data analysis, result in the collected data being representative of the students in the kindergarten class of 2010-11 who remain living in the United States. More information about the sample design can be found in the base-year User's Manual (NCES 2015-074).

[3] A base-year respondent has child data (scoreable assessment data or height or weight measurements, or was excluded from assessment due to lack of accommodation for a disability) or parent interview data from at least one round of data collection in the base 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 are 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 have been 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, and their schools. In kindergarten, the study also collected information from children's before- and after-school providers.

Students were directly assessed in each round of the first- and second-grade years. The untimed assessments were administered to the sampled students, one-on-one, by a trained assessor. The students were assessed in both fall and spring in reading, mathematics, and science, as well as executive function.[4] In the first-grade data collection, as in kindergarten, an assessment of Spanish early reading skills (SERS) for Spanish-speaking students who did not achieve a minimum score on an English language screener was included. The SERS was not included in the second-grade collections because most students passed the language screener by the end of first grade. In addition to the cognitive components, the direct assessments included measurements of the height and weight for all students in the fall and spring of each year.

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 first- and second-grade data collection rounds using computer-assisted interviews (CAIs). Most parent interviews were conducted by telephone.[5] In the fall of 2011 and fall of 2012, parents provided information about various educational and enrichment activities the child participated in during the previous summer. In these two data collections, questions were asked about educational activities in the home, use of a computer for educational purposes, reading books from summer book lists provided by the school, going to the library or bookstore, playing outside, outings, camps, summer school, tutoring, therapy services or special education programs, child health, hours spent watching television and playing video games, and nonparental child care. In the data collections in the spring of 2012 and 2013, the parent interviews asked about family structure, family literacy practices, parental involvement in school,

[4] 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.

[5] 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.

nonparental care arrangements, household composition, family income, parent education level, social support, parenting stress, and other demographic information. Parents were also asked to report on their children’s health, socioemotional well-being, peer victimization, 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, training, and experience; their classroom organization and resources; their curricular focus; their teaching, evaluation, and grading practices; and the level of involvement of students’ parents. A small number of items were included on the availability of computers and other electronic devices and on practices and professional development related to any Response to Intervention programs that may have been implemented in the teachers’ schools. Teachers 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 critical cognitive and noncognitive dimensions. Information was collected from general classroom teachers via self-administered paper questionnaires during both the fall and spring first- and second-grade data collection rounds.

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

School administrators were asked to provide information on the physical, organizational, and fiscal characteristics of their schools and resources; the schools’ learning environment and programs; the communities surrounding the schools; characteristics of the schools’ staff members; and their own background and experience. School administrators were also asked whether they spoke a language or languages other than English with students and families and about the implementation of practices related to Response to Intervention programs. Information was collected from school administrators via self-administered paper questionnaires during the spring first-grade and spring second-grade data collections.

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 2011 and spring 2012 first-grade and fall 2012 and spring 2013 second-grade rounds of data collection. Because the direct assessment scores were recalibrated at the end of each year of data collection, information about the recalibrated kindergarten scores is also included. Chapter 2 provides details about the design of the direct child cognitive assessment battery for the ECLS-K:2011 kindergarten, first-grade, and second-grade rounds. 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, describing the field test item pool, item analysis, and results, and the development of the final assessment forms used in national data collection. It also includes a description of item quality and reliability. Chapter 5 describes the psychometric characteristics of the direct cognitive assessment battery used in national data collection, including the approach to and types of scoring, choosing the appropriate scores for analysis, and measuring gains, or growth in cognitive knowledge and skills over time. Chapter 6 describes the psychometric characteristics of the executive function measures for first and second grade. Chapter 7 provides information on the psychometric characteristics of the indirect measures, including the *Children's Behavior Questionnaire (CBQ),* the *Temperament in Middle Childhood Questionnaire (TMCQ)*, 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 four appendixes supplementing the information in the main text. Appendix A presents the plots of IRT test information functions in reading, mathematics, and science. Appendix B provides details on how the students' responses to assessment items were prepared for and used in PARSCALE, the computer program used for estimating item response theory (IRT) models from which assessment scores were produced, as well as what quality control checks were performed on the assessment data. Appendix C lists the ECLS-K:2011 kindergarten, first-grade, and second-grade reading, mathematics, and science IRT item parameters. Appendix D presents the plots of IRT theta and scale score distributions in reading, mathematics, and science.

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, there is overlap in the content and actual items included in the assessments of the ECLS-K:2011 and the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) for comparability in the knowledge and skills measured in each study. However, despite this overlap, the direct cognitive assessment scores released for the ECLS-K:2011 in the kindergarten, kindergarten through first grade, and kindergarten through second grade data files are not directly comparable to those of the ECLS-K.[1] 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 goal of minimizing development and administration time and costs as well as the burden on students and teachers affected the structure of the assessments and the kinds of assessment items that could be used. 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. The experts recommended that the knowledge and skills assessed during each

[1] ECLS-K:2011 Restricted-Use Kindergarten Data File and Electronic Codebook (NCES 2013-060), ECLS-K:2011 Restricted-Use Kindergarten –First Grade Data File and Electronic Codebook (NCES 2015-070), ECLS-K:2011 Public-Use Kindergarten-First Grade Data File and Electronic Codebook (NCES 2015-086), ECLS-K:2011 Restricted-Use Kindergarten-Second Grade Data File and Electronic Codebook (NCES 2015-050), and ECLS-K:2011 Public-Use Kindergarten-Second Grade Data File and Electronic Codebook (NCES 2017-286).

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.

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.[2] 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,[3] current curriculum standards for kindergarten through third grade from California, Florida, New Jersey, Texas, and Virginia[4] 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 reading skills, vocabulary, and reading comprehension.

[2] Although this report focuses on the first-grade and second-grade 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 was targeted in the prior rounds and what content is to be expected in the subsequent rounds, as well as how the first-grade and second-grade content relates both to the overall framework and to the content measured in earlier and later rounds.

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

[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.

2.1.1.1 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.

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 further develop 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 (as defined in 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).

In the first-grade administrations, all of the skills listed were assessed; in the second-grade administrations, however, only the more advanced skills of blending and substitution of sounds were assessed.

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 and the National Association for the Education of Young Children 2008). The ECLS-K:2011 first-grade 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. Familiarity with print items were relatively easy and not appropriate for inclusion in the second-grade assessment.

Recognition of letters and sounds connects spoken language to written language. This is one of the first skills in early reading (International Reading Association and National Association for the Education of Young Children 2008). ECLS-K:2011 assessment items related to letter and sound recognition ask children to perform tasks such as choosing a specific letter from a set or giving 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. Several items requiring recognition of letters and sounds were included in the first-grade assessment. Only a single letter-sound item was included in the second-grade assessment, as these item types are too easy at this grade level.

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). Sight word items of varying difficulty were included in both the first- and second-grade assessments.

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. The first- and second-grade assessments included receptive vocabulary items assessing words used in the context of a sentence or paragraph. There were no items measuring expressive vocabulary in the assessment for either grade.

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 children’s 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 texts. For example, a child is asked to list the three things that fell out of the backpack of a girl featured in a story.

- **Integrate/interpret.** Test questions in this category ask readers to make complex inferences within 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 throughout the elementary school years. 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 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 are literary texts and one-third of the passages are informational texts.

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

To enable cross-cohort comparisons, which is one of the goals of the ECLS-K:2011, continuity between the ECLS-K (1998–99) 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. 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 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. And the **basic skills** and **vocabulary** categories are similar in both the ECLS-K and the ECLS-K:2011 frameworks.

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 design. 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 the framework 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 that, 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); with the National Council of Teachers of Mathematics *Principles and Standards for School Mathematics* (2000); and with state standards of California, New Jersey, Tennessee, Texas, and Virginia.[5] 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 following content categories: **number properties and operations**, **measurement, geometry, data analysis and probability,** and **algebra**.

[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.

2.1.2.1 Number Properties and Operations

From kindergarten through third grade, 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 to find sums or differences using numbers less than 20 when given concrete models or pictures. As children advance in age and grade, they are 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 are required to move from concrete representations of operations and ideas to more abstract representations and algorithms. In the first-grade and second-grade assessments, this content category is measured with questions asking children to identify certain single-digit and two-digit numbers; addition, subtraction, multiplication, and division problems with two, three, and/or four digits; and items assessing knowledge of relative quantity (e.g., fraction comparisons).

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 first-grade and second-grade assessments, this content category is measured with questions asking children to estimate the weight of an object on a scale, determine distance traveled in a certain time, and 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 first-grade and second-grade assessments, this content category is measured with questions asking children to identify basic shapes and to combine basic shapes into other 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 first-grade and second-grade assessments, this content category is measured with questions asking children to read basic graphs, indicate the probability of coins landing heads up, and relate fractions to probabilities.

2.1.2.5 Algebra

Algebra refers to the techniques of identifying solutions to equations with one or more missing pieces or variables, and also includes knowledge and understanding of patterns. Specifically, children are evaluated on their ability to recognize, create, explain, generalize, and extend patterns and sequences. In the first-grade and second-grade assessments, 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)[6] were analyzed to find a commonality of topics that are taught at the lower grade levels assessed in the ECLS-K:2011. In these states and for each grade level, three or four standards were specified for each of four common reporting categories: **scientific inquiry, physical science, life science,** and **Earth and space science.** These four reporting categories were selected as the content categories for the ECLS-K:2011 science assessment framework.

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. Examples of items in this content category in the first-grade and second-grade assessments are items about a microscope, differences in the ways animals move, and interpretation of data from observations.

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 and observe differences in the way things move. In

[6] 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.

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 gases; and understand that energy comes from the Sun to the Earth in the form of light and heat. In the first-grade and second-grade assessments, this content category is measured with questions about energy, the materials from which common objects are made, states of matter, and changes in the state of matter, such as condensation.

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 first-grade and second-grade assessments, this content category is measured with questions related to what animals eat, animal adaptations, and the functions of the human body and parts of plants.

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 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 first-grade and second-grade assessments, this content category is measured with questions about weather, animal habitats, soil properties, and the solar system.

2.2 Item and Time Allocation Across and Within Subject Areas

For the first- and second-grade 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 (about 30 minutes) than for the mathematics (about 15 minutes) and science (about 15 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 falls 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.

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.[7] 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.

[7] Note that although the framework percentages (in reading, mathematics, or science) may remain the same across grade levels, the assessments do not. For example, although measurement items account for 5 percent of the overall items in the kindergarten through second grade assessments, in kindergarten the measurement items administered are, on average, less difficult than those administered in second grade.

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 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, because they were administered just a subset of all items developed for the specific grade, 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 specific constraints on content, and the feasibility of scaling without the necessity of a major field test with random exposure of items across groups. The reading and mathematics components of the ECLS-K:2011 kindergarten assessment battery were multistage adaptive tests in both rounds. In the spring of kindergarten, the science assessment was a single-stage test. In subsequent rounds, the science assessment was a multistage adaptive test. (The Spanish early reading skills (SERS) assessment was single-stage in the rounds in which it was administered.) 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[8] 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.

[8] For multiple-choice items, a child's actual selected response was entered into the CAPI program. For open-ended items, the assessor scored the item based on the item's rubric and then entered either a correct or incorrect code into the CAPI program.

Although the second-stage tests were 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 was essentially performing on grade level received items that spanned the curriculum for that child's grade. Children whose achievement was above or below grade level were given items with difficulty levels that matched their individual level of knowledge and skills at the time of testing rather than a grade-level standard. Children evaluated at ability levels much higher than average, as measured by the routing test, were given test items that were more difficult (including some above grade level), while children evaluated at ability levels well below grade level received a second-stage test with proportionately more easy items (including some below grade level).

As noted earlier, two cognitive assessments were not adaptive. The kindergarten 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 in the kindergarten year were not as diverse as originally anticipated; therefore, a single-stage science assessment was deemed adequate. More information on the design of the kindergarten science assessment can be found in the *ECLS-K:2011 Kindergarten Psychometric Report* (Najarian et al. 2018). A single-stage Spanish early reading skills (SERS) assessment[9] was administered in kindergarten and first grade 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 assessment and routing of children through the cognitive assessment battery can be found in sections 4.2 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 in the

[9] The SERS used a single-stage design because the assessment was relatively short and included only items in the easier content categories. An adaptive design, which is used to more efficiently assess domains while limiting burden related to time of administration or difficulty, was not necessary.

kindergarten and first-grade rounds of data collection. By the spring of first grade, nearly all children demonstrated sufficient English proficiency to be assessed in English, so the Spanish-language assessments were not used in the second-grade rounds of data collection. More information about how children's home language affected the assessments children were administered in each round of data collection is provided in chapter 5.

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 administered in the kindergarten and first-grade rounds, and results from the screener were used to route children out of the cognitive assessments in English *after* being administered a subset of the items in the full reading assessment. 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 considered 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.

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 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.

In addition, Spanish-translated versions of the kindergarten and first-grade mathematics assessments were administered to those children administered the SERS. 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.

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 from 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 the PARSCALE computer program. 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 development of the longitudinal scales that are produced using assessment data from the kindergarten through second-grade data collection rounds, which allow for the measurement of gains in knowledge and skills across time. Evaluation of common items, the two calibration methods used (concurrent and chain-linking), and the computation of final scores are also discussed.

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 used to develop robust scales to facilitate score interpretation.

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.

Before IRT analysis was performed, 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 who had answered fewer than 10 questions in the assessment for the domain, and remove them from analyses.[1] Only items actually attempted by the child were counted toward the scoreability threshold. For the purpose of identifying unscoreable cases, codes for "don't know" were treated as incorrect responses for open-ended items. Conversely, codes for "don't know" for multiple-choice items were treated as invalid responses and were not included in the count of scoreable items. Before being removed from further 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.

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) (Lord and Novick 1968), was carried out separately for each round of data collection and for each subject area assessment using Educational Testing Service (ETS) proprietary software, F4STAT. Sets of statistics were produced for each item, as well as summary statistics for the router and each second-stage 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

[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.

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 (e.g., 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 unambiguously correct 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 the mean section score for test takers choosing incorrect responses, and the *r*-biserial should be positive, ideally .3 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. The *r*-biserials calculated during the classical item analysis for each assessment domain 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 chapter 4 for the routing and discontinue rules used in the assessments).

Summary statistics from the classical 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 .8 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 among those who were administered the second-stage forms relative to the variance in ability in 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 counts of the number of children administered each form (which were available in 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 assessment round and test form, 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 (Lord 1980) 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 his or her 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 examines 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 design characteristic assumption is that the test is not speeded, meaning that the positions of items relative to the beginning or end of the test do 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 (i.e., kindergarten and first grade). 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 all 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 in order to accommodate growth longitudinally, 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 that builds on skills mastered earlier.

3.2.1 Dichotomous Item Calibration

In the ECLS-K:2011 assessments, a dichotomous item is defined as an item for which the response is scored as either correct or incorrect; there is no partial credit given. In the case of a multiple-choice item, a correct response is scored if the correct response option is selected by the child; selection of a different response option would result in an incorrect score. For an open-ended item, the scoring rubric defines what responses are scored correct; responses not identified as acceptable correct responses in the scoring rubric are scored as incorrect. For example, if an item requires a two-part response, partial credit is not given if only one part of the response is given. The item would be scored correct only if the two parts of the response are provided, and scored incorrect if only one or neither of the two parts of the response are provided.

Figure 3-1 is an example of a graph of the logistic IRT function for a hypothetical dichotomous test item. The graph shows the most general model, the three-parameter (3PL) IRT model. The three item parameters are *a* (discrimination), *b* (difficulty), and *c* (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

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

where

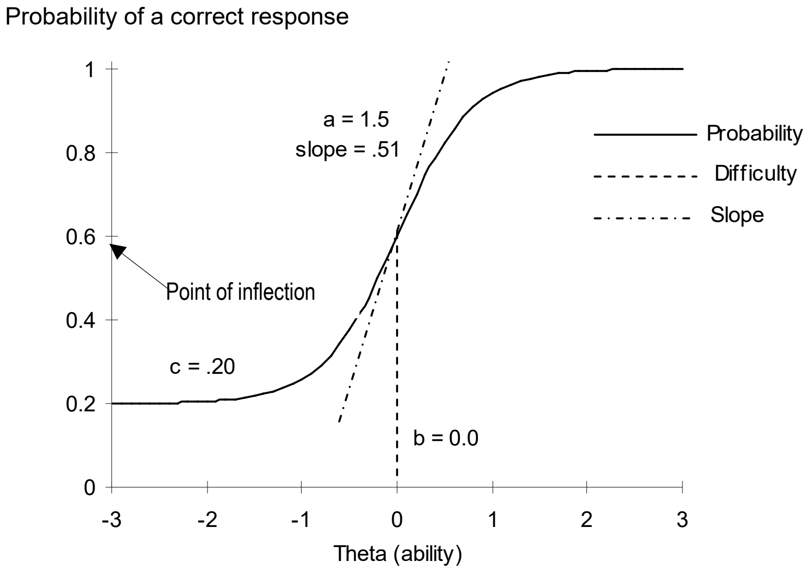
θ = ability of the test taker;

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 = 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 dichotomous test item



NOTE: *a* = parameter for discrimination; *b* = parameter for difficulty; and *c* = 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 2005-062), 2005.

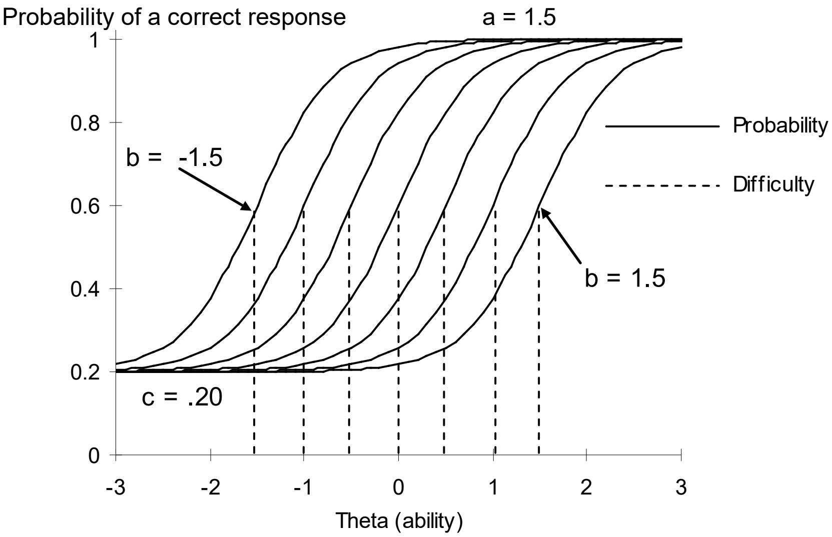
The IRT *c* parameter represents the probability that a test taker with very low ability will answer a multiple choice item correctly. In figure 3-1, out of 100 people with very low ability, 20 would get the item correct. Note that the *c* parameter does not necessarily equal 1 divided by the number of response options (e.g., .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), and extends from minus infinity (-∞) to infinity (∞). 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. The *b* parameter is identified at the point of inflection of the logistic function. The point of inflection occurs farther to the right for more difficult items and farther to the left for easier ones.

Figure 3-1 shows the logistic function for a dichotomous item having difficulty of $b = 0.0$, with the point of inflection at .6. This means that test takers with ability $\theta = 0.0$ have a 60 percent chance of getting the answer correct, or that out of 100 people with ability, or theta, equal to 0.0, 60 would be expected to answer the question correctly.

Figure 3-2 is an example of a graph of the logistic functions for seven hypothetical dichotomous 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 when $\theta > b$.

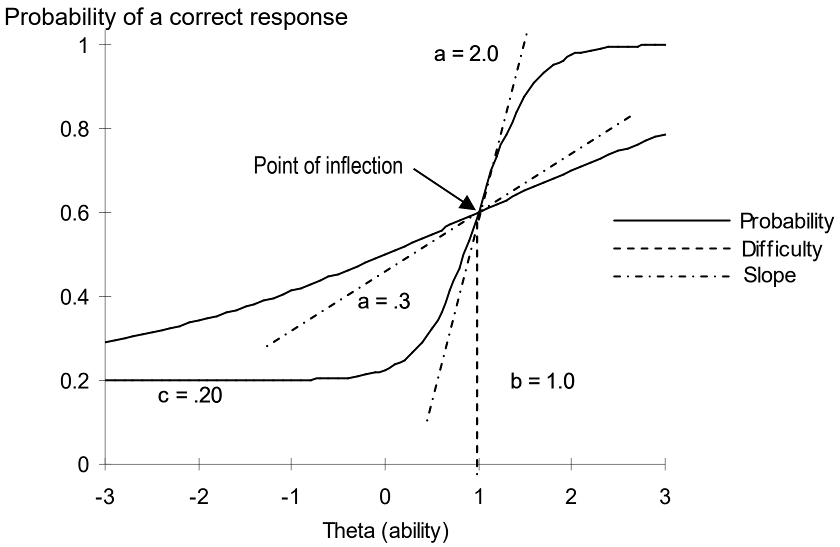
Figure 3-2. Three-parameter IRT logistic functions for seven hypothetical dichotomous test items with different difficulty (*b*)



NOTE: *a* = parameter for discrimination; *b* = parameter for difficulty; and *c* = 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 2005-062), 2005.

The discrimination parameter, *a*, is proportional to the slope of the logistic function at the point of inflection (slope=$0.425a_i(1-c_i)$). 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 dichotomous 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 balanced across the ability scale 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 dichotomous 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 2005-062*)*, 2005.

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 .25 or less for items with four response options, but they may vary with difficulty and, of course, the number of response 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 not estimated, can be used if the likelihood of guessing is very low. In a one-parameter IRT model, i.e., Rasch model, items are assumed to discriminate equally well, and the *c* parameter is not estimated. Certain tests can be performed on the data to determine which IRT model (a one-, two-, or three-parameter model) fits the data best (see section 3.2.4).

3.2.2 Polytomous Item Calibration

A change in scoring methodology for some items was implemented after administration of the first-grade assessments in the ECLS-K:2011. For all rounds of data collection in the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) and for the first release of scores from the kindergarten rounds of the ECLS-K:2011, dichotomous calibration was used for all assessment items. When the first-grade data were being processed, the decision was made to use polytomous calibration for the items associated with reading passages, and for groups of items that represent testlets, or small groups of items all measuring the same skill.

As stated above, one of the tenets of IRT modeling is the assumption of local item independence.[2] When multiple items are associated with a common reading passage or when similar items are presented in groups, or testlets (for example, when the correct response for a mathematics computation item is necessary to correctly respond to a subsequent mathematics item), individual items within a set are not necessarily, nor likely, to be conditionally independent, which means that the assumption of local item independency may be violated. For this reason, an analysis to check for LID was conducted by comparing dichotomous and partial credit polytomous scoring models for the kindergarten reading, mathematics, and science data.

In a dichotomous scoring model, for an item set containing four items associated with the same reading passage, each item would be calibrated individually, resulting in four separate item scores with possible values of 0 for an incorrect response or 1 for a correct response. In a polytomous scoring model, all items associated with a passage or that measure the same skill are combined into a single, polytomous item set. For example, for the same set of four items above, associated with the same passage, the item responses would be combined into a single item with a possible score of 0, 1, 2, 3, or 4, equaling the number of items in the set that were responded to correctly (e.g., if correct responses for three of the four items were given by the child, the child’s score would be a 3 for that polytomous item.)

Figure 3-4 contains a plot of the theoretical item characteristic curves for a polytomous item from a testlet including two individual items. The horizontal axis represents the theta scale, while the vertical axis represents the probability of a correct response, similar to the plot in figure 3-1 for a dichotomous item. The lines in figure 3-4 represent the theoretical item characteristic curves based on the

[2] High levels of dependency result in biased item parameter estimates, and underestimation of low thetas and overestimation of high thetas. The distribution of theta is spread out at the tails and flattened in the center due to the greater number of low and high thetas.

item parameter estimates and the equation for the generalized partial credit IRT model. The fundamental equation of this model is the probability that a person with an estimate θ will have, for the ith item, a response that is scored in the jth of *m* ordered score categories, defining the probability of a correct response P$_{ij}$:

$$P_{ij}(\theta) = \frac{\exp[\sum_{v=1}^{j} a_i(\theta - b_i + d_v)]}{\sum_{c=1}^{m} \exp[\sum_{v=1}^{c} a_i(\theta - b_i + d_v)]},$$

where

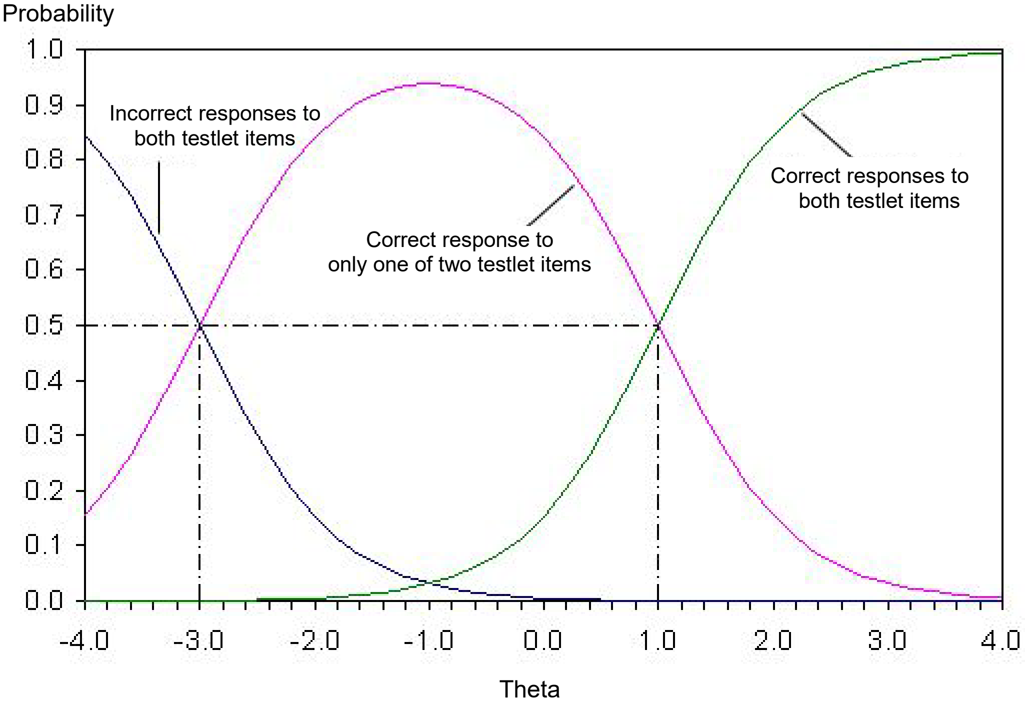
m_j = number of items in the testlet or associated with the same passage in polytomous item j;

a_i = discrimination of item i;

b_i = difficulty of item i; and

d_v = category j threshold parameter for item i.

Figure 3-4. Three-parameter IRT logistic function for a hypothetical polytomous test item of a two-item testlet



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 and 2001 Assessment.

The IRT *a* and *b* parameters define the discrimination and difficulty, respectively, in both the dichotomous and polytomous item calibration models. (The *a* parameter for polytomous items represents the *same* magnitude of the maximum slope for *each* of the curves.) For polytomous items, children are unlikely to guess the correct responses for all items, thus, $c = 0$. The category threshold parameter, *d*, is specific to the polytomous model only, and is interpreted as the relative difficulty of step j in comparing other steps within the polytomous item; or the probability of obtaining a score on an item that is above the threshold. Because the threshold is a negative function of the item-category parameter, a more difficult item category has a higher threshold value. In figure 3-4, $d_1 = 2.0$ and $d_2 = -2.0$. The place on the theta scale where students are equally likely to have incorrect responses to both items or a correct response to one item is -3.0 ($b - d_1 = -1.0 -2.0$), and the place on the theta scale where children are equally likely to have a correct response to one item or correct responses to both items is 1.0 ($b - d_2 = -1.0 + 2.0$). Children low on the scale are expected to have incorrect responses, children higher on the scale are more likely to have part of the set of items correct, and children highest on the scale are most likely to have all items correct.

The results comparing the dichotomous and partial credit scoring models for the kindergarten reading data showed that the scores correlated very highly (.99), indicating that the dichotomous scoring model was appropriate when just kindergarten data were considered. When all the items are treated dichotomously, the issue of local item dependence is ignored, which when looked at in a cross-sectional way is not so consequential. However, for longitudinal measurement and vertical scaling, research has shown that the standard errors of estimates of ability may become inaccurate in later grades when the testlets and item sets associated with the same passage become more complex and possibly less independent, and such inaccuracies can lead to poor estimation of the variability of scores, a result of violating the assumption of local independence. For this reason, it was decided to use a polytomous model when scoring the data once data from later rounds were available. Thus, scoring sets of items associated with passages and testlets as polytomous items should improve test equating over the time period of the ECLS-K:2011. In order for the scores to be comparable across round and grades, they must be computed using the same scoring model. Therefore, although the kindergarten scores correlate highly across the models in the kindergarten round, the kindergarten scores were recomputed using a polytomous scoring model to allow for comparisons across time and measurement of growth.

Similar to the reading data results, correlations between the dichotomous and partial credit scoring models were very high for both the mathematics and science data. However, testlets in these domains were not expected to become less independent with longitudinal progression, due to the variability in item content and presentation. Therefore, it was decided to retain the less complex dichotomous model for the kindergarten and subsequent data collections in mathematics and science.

3.2.3 Calculation of IRT Scale Scores

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*, *c*, and for the polytomous sets, the *d* 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.

3.2.4 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 assessments. The appropriateness of both the two-parameter IRT model and the three-parameter model was first investigated for the ECLS-K:2011 kindergarten assessment data.

Initially, one-, two-, and three-parameter IRT models were compared in terms of relative model fit (AIC [Akaike information criterion] and BIC [Bayesian information criterion]) using IRT software described in Haberman (2013). The three-parameter model with constant guessing was used, because the full 3-parameter model showed model identification issues in this software. The results of the comparisons are shown in table 3-1. The three-parameter model with constant guessing showed the best relative fit (lowest AIC and BIC), followed by the two-parameter model and then the one-parameter model. However, the difference in fit between the one- and two-parameter model was much larger than the difference between the two- and three-parameter model. As can be seen from the table, the results for both reading and mathematics are consistent.

Table 3-1. Comparison of one-, two-, and three-parameter IRT models of relative model fit (AIC and BIC)

Domain	Model	Sample	Items	Log-likelihood	Parameters	AIC	BIC
Reading	1 PL	34062	83	-809155.0	84	1618478	1619187
	2 PL	34062	83	-787672.6	166	1575677	1577078
	3 PL (constant guessing	34062	83	-785562.3	167	1571459	1572867
Mathematics	1 PL	33944	75	-686754.2	76	1373660	1374301
	2 PL	33944	75	1343928.6	150	1344229	1345493
	3 PL (constant guessing)	33944	75	1343287.9	151	1343590	1344863

NOTE: PL refers to parameter logistic model. AIC = Akaike information criterion. BIC = Bayesian information criterion.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

Other analysis included review of model fit, thetas, standard errors, and outliers, and was performed both on the data as a whole and for the items individually. Model fit was evaluated through visual inspection comparing the modeled item characteristic curve (section 3.2.5) and the actual data, as well as comparisons of the overall modeled and observed percent correct values. Items that were not modeled well as illustrated by the item characteristic curves, and those with modeled and observed percent correct values with discrepancies greater than 10 percent, were considered for removal from the analysis.[3] Theta estimates comparing the two-parameter and three-parameter IRT models were reviewed to determine if the thetas followed the expected trends (e.g., on average, thetas increase in subsequent data collections), and standard errors were reviewed to determine if the errors were higher with either model. Outliers were considered outside of five standard deviations of the theta mean, and were not observed in either model.

It was 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-

[3] In analyses of the national data, other factors, such as the item's content category, and where along the ability distribution the model discrepancies occurred, factored in to the decision to exclude an item from scaling and scoring.

parameter estimation resulted in a better fit for the multiple-choice items; therefore, the three-parameter model was selected for use throughout the ECLS-K:2011 study.[4]

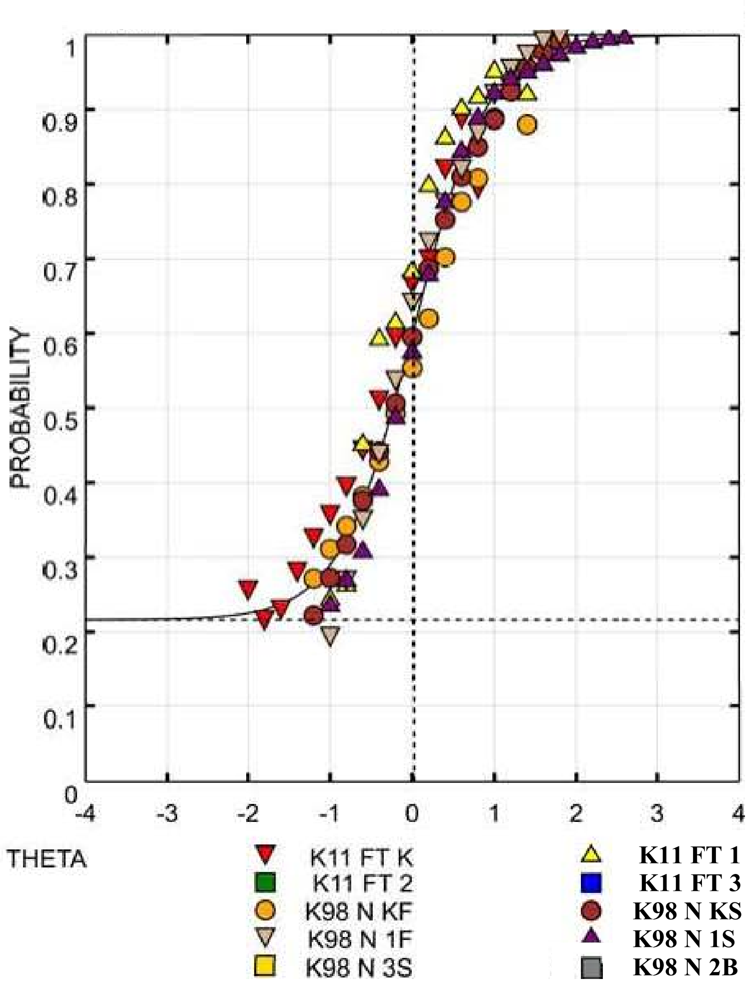
3.2.5 Evaluating Items Using Empirical Item Characteristic Curves (ICC)

As discussed above, the item parameter estimates can be used to graph the probabilities of correct answers across the entire ability range. 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 in addition to modeled data points is called an empirical ICC. The empirical ICC in figure 3-5 shows the fit of the three-parameter model to the actual data for a well-functioning dichotomous 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 3-6 shows the empirical ICC of a dichotomous item included in the assessment field test that did not function particularly well. Although about 37 percent of the test takers answered this item correctly, performance on this item was 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. Items such as this are excluded from consideration in subsequent assessment designs.

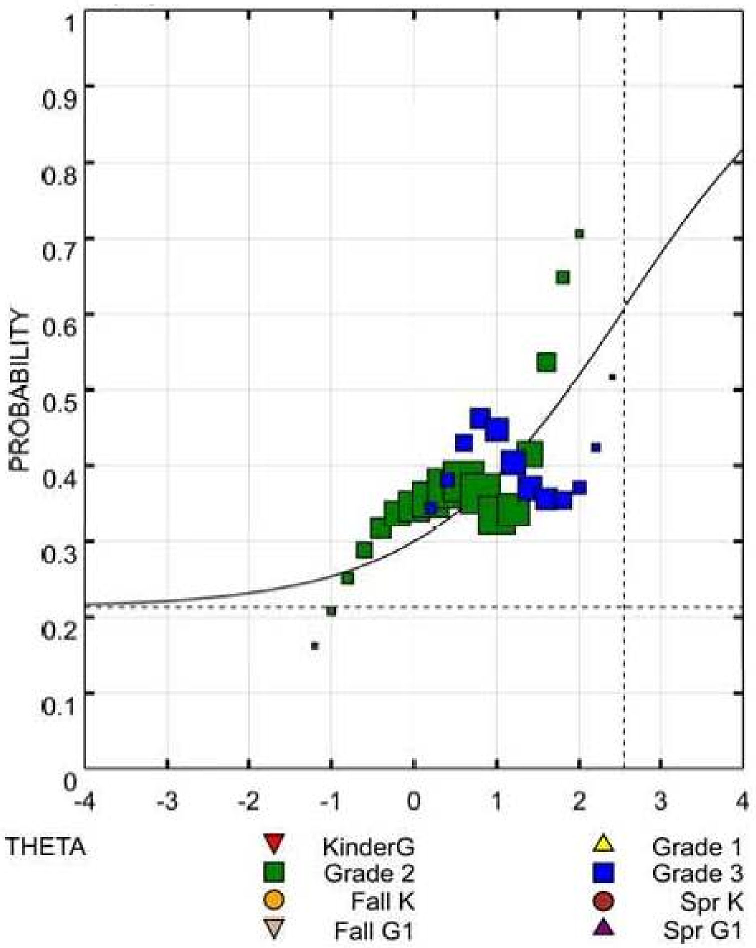
[4] The same model used in the first round of the study (the fall of kindergarten) must be used in subsequent rounds to permit longitudinal measurement across the entire study.

Figure 3-5. Example of an empirical dichotomous 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 3-6. Example of an empirical dichotomous 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.

3.2.6 Item Information and Measurement Precision

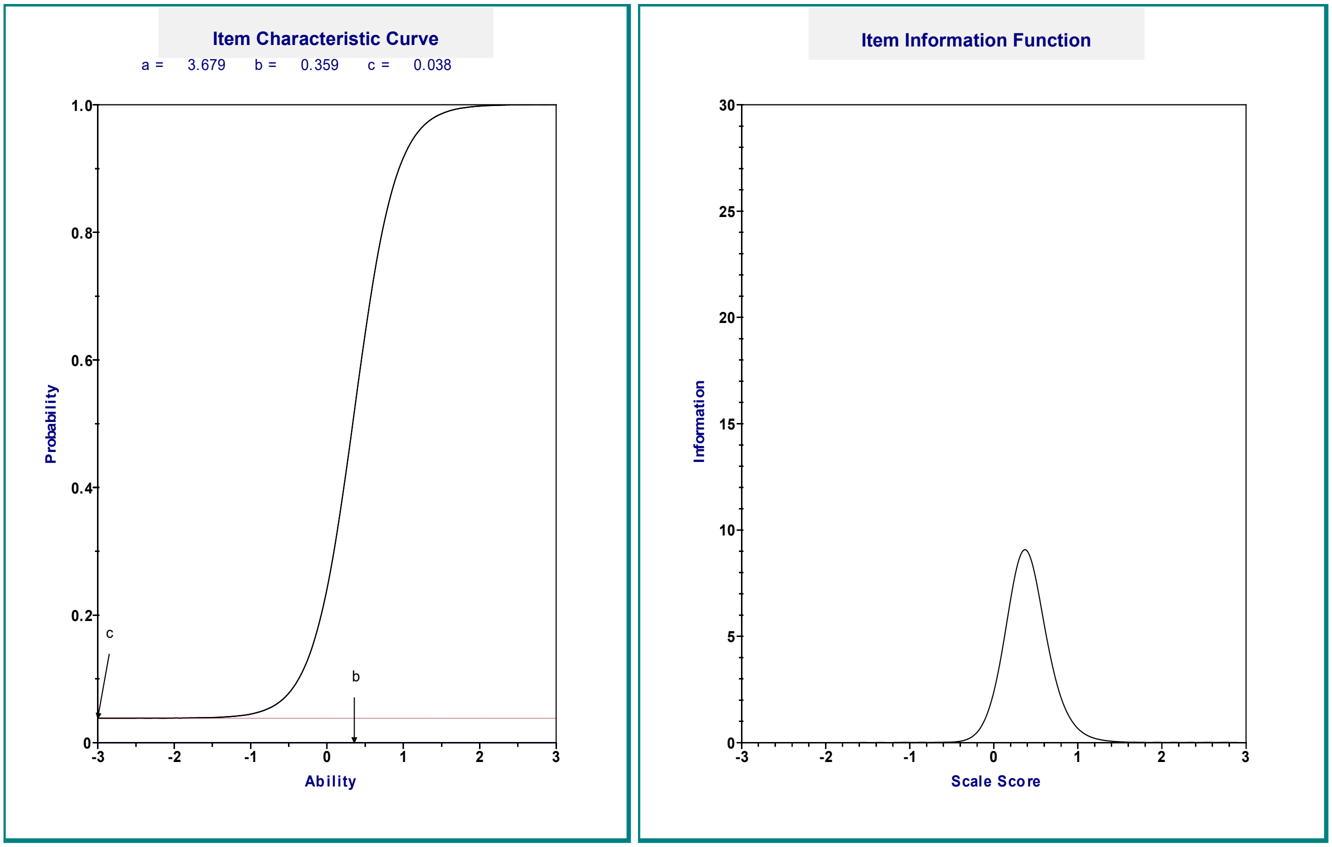
Another way to measure item and test quality is to estimate the item information function (IIF) and test information function (TIF). In psychometrics, the precision of parameter estimates at the various ability levels can be measured using the information function (Lord 1980). This is computed as a function of the reciprocal of the measurement error, denoted as σ^2. The information function (*I*) is defined as

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

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 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 ability level, a precise estimate of that test taker's theta, 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*, *c*, and as appropriate, *d* parameters, the IIF provides a visual representation of the place on the ability scale where the item measures best. Figure 3-7 shows the ICC and IIF for a hypothetical dichotomous item. This item has good discrimination and seems to measure well for test takers with a theta ability of approximately 0 to 1.

Figure 3-7. 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 for a dichotomous item i (I_i) is defined as

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

where

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

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

a = discrimination parameter;

c = guessing parameter;

θ = ability of the test taker;

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

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

The definition of the polytomous item information function for the three-parameter IRT model is a bit more complex, with a different formulation. For item i:

$$I_i\left(\theta\right) = a_i{}^2 \sum_{c=1}^{m} [T_c - \overline{T}_i(\theta)]^2 P_i(\theta),$$

where

$$P_i(\theta) \quad = \quad \frac{\exp[\sum_{v=1}^{j} a_i(\theta - b_i + d_v)]}{\sum_{c=1}^{m} \exp[\sum_{v=1}^{c} a_i(\theta - b_i + d_v)]};$$

$$\overline{T}_i(\theta) \quad = \quad \sum_{c=1}^{m} T_c P_{ic}(\theta)$$

θ = ability of the test taker;
j = jth of m levels of item i
m_j = number of items in the testlet or associated with the same passage in polytomous item j;
a_i = discrimination of item i;
b_i = difficulty of item i; and
d_v = category j threshold parameter for item i.

The aggregate of all the individual IIFs is the test information function, which is estimated using only the administered items with correct or incorrect responses. It is expected that an entire test measures ability more precisely than does a single item. 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 (TIF) will show high levels of information across the child's ability range.

The test information function (Lord 1980) 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

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

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.[5]

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:

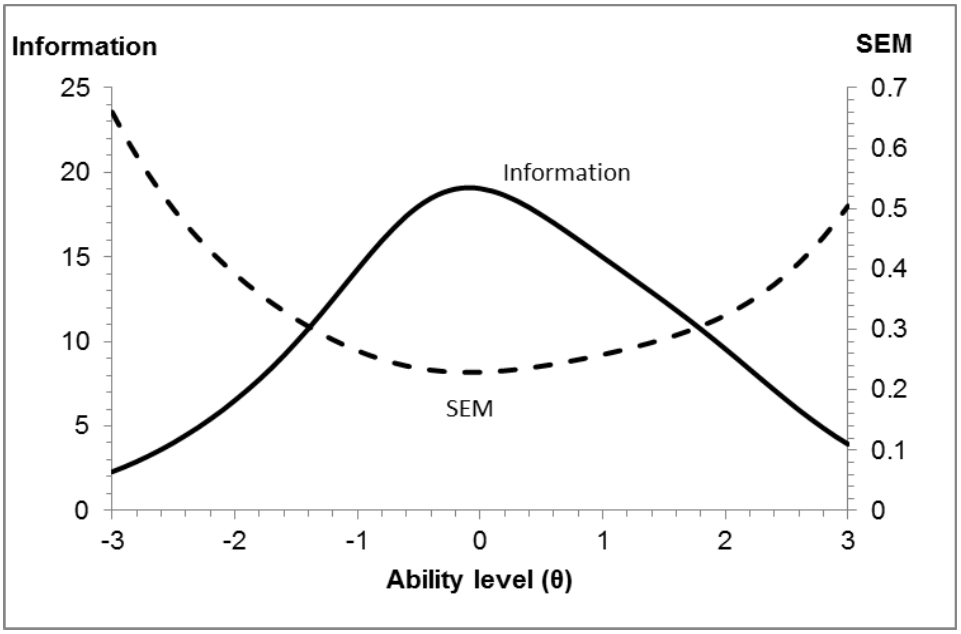
$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

An example TIF is shown in figure 3-8. 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. The *SEM* is conditional on ability; as the information increases, the standard error decreases.

The *SEM* (weighted) was estimated for each theta estimate for each domain in each assessment round. The *SEM* estimate can be acquired when each child's theta estimate is known. These estimated standard errors are provided in the data file for each of the thetas.

[5] See appendix A for test information plots. Item information plots are available upon request (ecls@ed.gov).

Figure 3-8. 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.

3.2.7 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.[6] 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, c*, and, if appropriate, *d* parameters, for each test item, iterating between the item parameters and ability estimates 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 (200). The convergence criterion and maximum number of cycles are

[6] The version of PARSCALE used to calculate scores for the ECLS-K:2011 is a customized version originally developed for the National Assessment of Educational Progress (NAEP), not the version that is currently commercially available.

based on guidelines adopted from NAEP. 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.[7] 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.[8] That is, after all of the kindergarten through second-grade assessments were completed there were six defined subpopulations of different ability levels related to time of testing (i.e., data collection round). The level of performance at each subsequent data collection round is, on average, greater than the levels of performance in the prior rounds. For example, the spring kindergarten subpopulation has, on average, a higher expected level of performance than that found for the same children during the fall kindergarten data collection. Similarly, the level of performance in the fall of first grade is, on average, greater than the levels of performance in the fall and spring of kindergarten but lower than the level of performance in the spring of first grade or fall of second grade.

For each grade, the identical assessment was used in both the fall and spring data collections, and the assessment included items measuring knowledge and skills the students might be expected to learn during the course of the school year. At the time of the fall assessments in each grade, only a few children may be able to answer the more challenging questions on the high second-stage form. Thus, there may be a limited set of data collected on the most difficult items in each domain during the fall rounds. 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 in each grade. As a result, when the two data rounds (fall and spring) at each grade 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 PARSCALE and other approaches to IRT is that they can incorporate prior information about the ability distribution (i.e., data from the previous round) into 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

[7] The empirical item characteristic curves noted above are a visual representation of the data and model fit.

[8] As used here, "subpopulation" refers to the data available at a point in time or around a given ability level (e.g., the fall or spring of kindergarten, the fall or spring of first grade, or the fall or spring of second grade). In IRT, subpopulations are created by dividing 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.

recalibrating all of the item parameters using Bayesian priors[9] 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.

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 (a, b, c, or d), 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 within a grade (fall and spring) is treated as a separate subpopulation with an independent ability distribution. The amount of shrinkage, or regression, toward the subpopulation mean is a function of the distance between the extreme value and its subpopulation mean 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

[9] A prior as used here is a proficiency (i.e., ability) distribution defined a priori to reflect prior expectations of the true distribution. In this case, the proficiency distribution is expected to be standard normal; thus, the prior is a standard normal distribution. PARSCALE uses separate prior (normal) distributions of ability for each subpopulation and optionally updates these priors with the estimated posterior distributions after each calibration iteration.

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 floor and ceiling 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 first-grade science assessment is designed to measure 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. Principal component analyses were run using Pearson correlations as a basic check on dimensionality of the data collected during the national administrations. 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 component. 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 content area experts 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 (Reise, Horan, and Blanchard 2011). 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 item content may vary within a component (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. Section 5.2.1 includes a discussion of the component analyses run in each domain for each assessment year to determine if each assessment in each domain was measuring a single, dominant component.

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 a measure of overall ability or performance (e.g., Black and White children with the same approximate theta estimate 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 kindergarten or first-grade 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 one language group compared to the other.

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 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.[10] Accordingly, 41 two-way tables were produced for each item, one for each theta interval. The first dimension of each of the 41 two-

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

way tables is population subgroup (e.g., White children versus Black children), and the other dimension is whether or not the child answered an item correctly. 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, or an 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 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 D-DIF for those items in the extremes of the distribution: the P-DIF procedure looks at differences in adjusted proportions of correct item responses, while the M-H procedure looks at the log odds ratios. For this reason, the M-H procedure is more susceptible than the P-DIF procedure 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 -.05 and +.05 are considered negligible. Values between -.10 and -.05 and between +.05 and +.10 are inspected to ensure that no possible effect is overlooked. Items with values outside the -.10 to +.10 range are more unusual and are identified as exhibiting DIF with practical significance (Dorans and Holland 1992).

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 -.10 or greater than +.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 since, as noted above, the M-H procedure is more susceptible than the P-DIF procedure to a false indication of C-level DIF for items at the extreme values of the difficulty distribution.

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 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. Individual items were checked for differential functioning using child's sex and race/ethnicity, and round of administration within each grade as analysis characteristics. The sex contrast compared males (reference group) with females (focal group). The race/ethnicity contrast groups included non-Hispanic 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 Development of the Kindergarten Through Second-Grade Longitudinal Scale

The study of the relationships between children's early childhood experiences at kindergarten entry and their gains in academic skills in subsequent grades required the development of a vertical assessment scale spanning kindergarten through second grade that had optimal measurement properties throughout the achievement range. That is, the assessments administered in each round (the fall and spring of kindergarten, the fall and spring of first grade, and the fall and spring of second grade) together needed to reflect the core curriculum elements covered in each domain at each grade level, and scores from each round needed to be comparable to one another. It was possible to meet these two requirements by ensuring that the test forms for a given domain that were administered in different rounds had common items and that there was overlap in the difficulty distributions of the items included in adjacent forms (e.g., the first-grade low and the first-grade middle second-stage math forms) and rounds.

Modeled after the general design in NAEP, at least 50 percent of the items should overlap across forms within grade and across the router and second-stage forms in adjacent rounds. Care was taken in the design to ensure that ample numbers of overlapping items across forms within grade and across assessments in adjacent grades were included to permit both horizontal (i.e., within round) and vertical (i.e., longitudinal) scales for each domain, beginning in fall kindergarten. Longitudinal measurement of science begins in the spring of kindergarten since a science assessment in the fall of kindergarten was not administered.[11]

3.5.1 Evaluating Common Item Functioning

Although the content and presentation of each of the common items were identical in each round in which the item was administered, it is still possible for the items to function differently. Since common items exist on adjacent second-stage test forms within grade and also across grades, some children are administered the same item on different forms in subsequent rounds, in a different item order, and/or 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, the difficulty of items in the

[11] The relatively short kindergarten single-stage science assessment (20 items) resulted in only about 40 percent of items overlapping between kindergarten and first grade, using the longer 43-item first-grade science assessment as the denominator.

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, data from each round within grade 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 were not common functioning would not be used as common items for the purposes of developing a vertical scale. These items are 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.

Assessing the common functioning of overlapping items in each domain across grades was done by comparing the *actual* performance on the common items with performance *predicted* by the IRT item and ability parameters, in order to identify discrepancies that would indicate differential functioning for any items. The comparisons of observed vs. predicted percentage correct for each question are based on the data for children who answered each of the items in each round of data collection, because the comparisons can be carried out *only for children who answered the question*. Many questions appeared in only one or two second-stage forms within a grade or after a discontinue point in a form. Thus most of the items were answered by only a subset of children tested in each round.

The results from the analysis of common item functioning do not represent the difficulty of the items, but rather the fit of the IRT model to the data, evaluated on the basis of comparisons of actual and predicted responses for all items answered. Little to no difference between the observed and predicted percent correct indicates common functioning of the items across time periods and good fit to the IRT model. Results of the DIF analyses conducted to examine common item functioning for the kindergarten, first-grade, and second-grade assessment items are included in section 5.2.2.

3.5.2 Concurrent Calibration and Chain Linking

The development of the longitudinal scales necessary for measuring gain over time begins by pooling the data from the fall and spring rounds within a grade level, as described above, and concurrently calibrating the data within grade. Then, the subsequent link between the assessments used at different grade levels relies on the presence of common items shared across the tests and was developed using chain-linking methodology. The combination of concurrent calibration *within* grade, and chain-linking *across* grades, results in a single, longitudinal scale.[12]

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 approach, 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 3.2.7, is particularly important for a longitudinal study, where the focus is on measuring gain and it is important to avoid floor and ceiling effects.

A chain-linking approach is used to place item parameters and ability (theta) estimates on the same scale *across* grades. TBLT[13] is an ETS-proprietary computer program that is used to estimate the linear relationship between two independently calibrated IRT scales. This is accomplished by minimizing the average squared difference between two test information functions estimated from the items common to both scales.

If the latent variable on the reference scale is denoted by θ_R and the latent variable on the new scale by θ_N, the goal is to find a linear transformation such that $\theta_R = A\theta_N + B$, where *A* and *B* are constants of the linear transformation of scale. For the three-parameter logistic model, the corresponding transformations of the item parameters are

$$a_R = a_N / A$$
$$b_R = Ab_N + B$$
$$c_R = c_N$$

The result of these transformations would be to keep the item response functions identical:

$$P_R(\theta_R) = P_N(\theta_N).$$

[12] After the data were cleaned as described above, preliminary item parameters were estimated using IRT procedures. These preliminary estimates were used in analyses examining item DIF and common functioning. Once these analyses were completed, final parameter estimates were calculated using concurrent calibration and chain linking for the set of items retained within each domain.

[13] TBLT stands for The B-Less Transformation since it replaces methods of computing the linear transformation parameters based on using only the "b" parameters.

The TBLT approach to estimating *A* and *B* described in Stocking and Lord (1983) tries to match test information functions for a set of common (or anchor) items.

The values are chosen to have a desired distribution (uniform, normal, etc.) over a desired range. This is a nonlinear least squares fitting problem that is solved iteratively in the program.

The estimated values of *A* and *B* are then applied to transform the parameter estimates for all the items in the new calibration. The success of the transformation is evaluated by comparing the fitted test characteristic curves as well as the fitted item response functions for the common items. The reference scale is thus transformed into the new scale, establishing a common reporting metric whereby scores for all rounds are comparable.

Backward chain-linking (i.e., linking assessment data from later rounds back to the assessment scale established for an earlier round), as opposed to forward chain-linking (i.e., linking assessment data from prior rounds to the current round) was selected for the development of assessment scores allowing for longitudinal analysis primarily to keep the theta scale fixed across the entire study. With backward chain-linking, the theta estimates from prior rounds remain the same, even once data from subsequent rounds are collected. Thus, the thetas estimated for the kindergarten rounds remain the same as originally reported once the first-grade thetas are computed and reported, and the thetas estimated for the kindergarten and first-grade rounds remain the same as originally reported once the second-grade thetas are computed and reported.[14] For the ECLS-K:2011, multiple transformations were required to place all item parameters and child ability estimates on the kindergarten scale. For first grade the fall and spring first-grade data were first concurrently calibrated and then were transformed, or converted, to the kindergarten scale using chain linking. Similarly, for second grade the fall and spring second-grade data were concurrently calibrated and then, were (1) first transformed onto the first-grade scale and then (2) transformed onto the kindergarten scale using chain linking. Thus, kindergarten, first-grade, and second-grade data are comparable as a result of the backward chain-linking methodology used.

The use of the chain-linking methodology to link scores across grades is a change from the methodology used in the ECLS-K. In the ECLS-K, the longitudinal scales necessary for measuring gain over time were developed by pooling (i.e., concurrently calibrating) the four rounds of kindergarten and first-grade data with the data from the ECLS-K third, fifth, and eighth grades. All data (separately by

[14] Please note that the kindergarten thetas were re-estimated using the polytomous calibration described above, and are thus on a different scale in rounds subsequent to the base year. More details can be found in section 5.2.4.1.

domain) were pooled and a single, concurrent calibration was run, maintaining each round of data as a separate population. Thus, the concurrent calibration spanned the entire kindergarten through eighth-grade data collection waves. During initial analyses of the ECLS-K:2011 data, it was suggested that, based on research conducted since the inception of the ECLS-K, the ECLS-K:2011 calibration be changed to a hybrid using concurrent calibration within grade and chain-linking methodology across grades.

Research comparing model fit of the data from completely concurrent (i.e., all rounds and data pooled into a single concurrent calibration) and completely separate chain-link (i.e., data from each round calibrated separately) scaling methodologies has shown somewhat mixed results with a slight advantage to the concurrent procedures. For example, Tsai et al. (2001), Hansen and Beguin (2002), Beguin and Hansen (2001), and Meng (2007) found small but consistent effects favoring concurrent procedures over completely separate chain-linking procedures. However, while Meng found that a completely concurrent approach was superior to a completely separate chain-linking approach, hybrid approaches (i.e., those using both concurrent and chain-linking methodologies) were superior to the other two “pure” approaches. Meng was carrying out calibrations spanning six grade levels. Thus, use of a hybrid model seems appropriate when developing scores for longitudinal measurement across that many grades, as will be done by the end of the ECLS-K:2011 (which is following children from kindergarten in the 2010-11 school year through the spring of 2016).

3.5.3 Computation of Final Scale Scores

IRT-based scale scores are derived from the IRT item parameter estimates and ability estimates (θ). 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. These probabilities are then summed over all unique items in each domain, for all rounds. For example, a child who is tested at all rounds through the spring of second grade (fall kindergarten, spring kindergarten, fall first grade, spring first grade, fall second grade, and spring second grade) will have six ability estimates and the associated scores for each round.

Further information on the psychometric characteristics of the final scores is provided in chapter 5.

This page intentionally left blank.

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

The previous chapters described the framework and general methodologies that guided the development of the ECLS-K:2011 assessments. This chapter provides information about the development and actual construction 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 first-grade and second-grade data collections. The executive function assessments are discussed in chapter 6; the study administered existing, well-tested executive function assessments so no development or field testing for items in this domain were needed.

4.1 Development of the Item Pool

In the first step of the process to develop the pool of items that could be used in the ECLS-K:2011 assessments, 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 related to some topics covered 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. Additionally, new kindergarten and first-grade 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, or SERS, assessment), were developed for the ECLS-K:2011. Once the initial pool of items was developed for the reading assessment in English,

[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.

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.

4.1.1 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.

4.1.2 Methods Used to Analyze Data for Design of the National Assessments

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.

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] in addition to items newly developed for the ECLS-K:2011 to measure concepts not included in the earlier studies. 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 the quality of the items newly developed for the ECLS-K:2011 and how the items that were previously developed for earlier studies 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 functioning.

In order to measure each child's status accurately in the national assessment, it is important that each child receive 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 multiple sources. The sources listed below were used in the design of the first-grade national assessments. Note that 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 ECLS-K K-1 general knowledge assessment were used.

- 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);

[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.

- 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).

For the second-grade national assessment design, data from the ECLS-K:2011 kindergarten and first-grade national assessments were also used when available to inform the design. Kindergarten data from the ECLS-K:2011 2009 field test and the ECLS-K fall and spring kindergarten national assessments were not used for two main reasons: (1) the ability estimates for children from the ECLS-K:2011 and ECLS-K kindergarten rounds were well below those for the target sample of second-graders in the ECLS-K:2011; and (2) item parameter estimates using the ECLS-K first-grade data, in conjunction with the available ECLS-K:2011 samples and data, provide adequate information to inform the second-grade design. Therefore, for the second-grade national assessments, the following data were pooled for the analysis:

- 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 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[6] (approximately 900 cases);
- ECLS-K spring third-grade national data collection (approximately 14,000 cases).
- ECLS-K:2011 fall kindergarten national data collection (approximately 16,000 cases);[7]

[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 the *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062) (Pollack et al. 2005).

[6] As noted earlier, 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 2005–062) (Pollack et al. 2005).

[7] A science assessment was administered beginning in spring kindergarten, so the fall data are available only in the reading and mathematics domains.

- ECLS-K:2011 spring kindergarten national data collection (approximately 18,000 cases); and
- ECLS-K:2011 fall first-grade national data collection (partial data available, approximately 4,000 cases).[8]

A separate IRT calibration focusing on the effects of language of administration at the first-grade level was conducted in concert with analysis of the kindergarten sample 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 the study did not plan to measure Spanish skills beyond those considered basic. The SERS items were not re-evaluated before the first-grade collections; the original analysis indicated the SERS set would be at the appropriate difficulty for English language learners (ELLs) in first grade.

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. Second, the IRT model used requires at least 400 examinees per item in order to obtain stable item parameter estimates. Pooling the field test data with the large samples from the ECLS-K:2011 and the ECLS-K national 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. 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 first or second grade) 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.

[8] Only partial data from the ECLS-K:2011 fall first-grade national data collection were available at the time of second-grade assessment design.

4.1.3 Criteria Guiding the Selection of Items for the National 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 assessments 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, and 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.
- **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.

The adaptive, two-stage 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 each of the ECLS-K:2011 data collection 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 assessment forms of each subject area at each round. The ability (theta) estimates for the ECLS-K:2011 assessment rounds 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 data collections.

The first-grade and second-grade design analyses using the pooled data showed differences in the estimated mean ability levels between the ECLS-K:2011 field test and the ECLS-K and the ECLS-K:2011 national samples at similar grade levels. Several factors may have contributed 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 representative of the population. Another possible factor could have been real changes in the kindergarten population in the interval between 1998 and 2009 with respect to prior exposure to early learning experiences. 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 ECLS-K:2011 national, ECLS-K national, and the ECLS-K:2011 field test distributions. This range, from roughly two standard deviations below the lowest estimated fall mean ability level to two standard deviations above the highest spring estimated mean ability level, was expected to include at least 95 percent of children in the ECLS-K:2011 national sample. Another reason for extending the difficulty range of the items at both the low and high end of the ability range is to avoid floor and ceiling effects in the national assessments.

The estimated 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 range of the items to the ability range 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 generally were not considered for inclusion in the national 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 each domain so that items could be selected for the routers and the three second-stage tests. For each estimated ability range, the low end of the range was computed using the mean ability level and the associated standard deviation of the lowest scoring sample, while the high end of the range was based on the mean ability level and the associated standard deviation of the highest scoring sample. Generally, the lowest ability level ranged from two standard deviations below the lowest mean to the highest mean; the middle ability level ranged from one standard deviation below the lowest mean to one standard deviation above the highest mean; and the highest ability level ranged from the lowest mean to two standard deviations above the highest mean. The router was designed to have items with difficulties spanning the entire expected range of ability, because having information about a 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 ensures 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 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 she was tired and not paying close attention to questions she 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 she received in the middle form might have been relatively easy for her.

The original design concept included the intention to re-use assessment forms with more difficult items added to cover the expected increase in ability levels in subsequent grades. For example, the kindergarten second-stage low, middle, and high forms were referred to as forms A, B, and C, respectively. For first grade, form A was to be dropped, forms B and C were to be used as the low and middle forms, and a new form D was to be added. Though some second-stage tests were to remain unchanged, it was expected that the routing form for each grade would need to be revised, with new cut score values to route to the second-stage forms.

However, as the cohort progressed from kindergarten through second grade, it became clear that this design concept could not be implemented in its entirety for multiple reasons. First and foremost, with the addition of data from the ECLS-K:2011 national assessments, improved estimates of child ability levels showed differential changes in the anticipated increases in average ability level with increasing grade level; thus, some items on the forms were too easy or too difficult for children at the next grade level. Second, as items that were newly developed for the ECLS-K:2011 were tested on a larger sample, item psychometrics showed that some of the items did not satisfy the minimum requirements to be included in subsequent assessments. Finally, some items were shown to simply require much more time to administer than originally anticipated and were removed to reduce assessment burden. Despite the fact that the original design could not be followed completely, the routing tests were revised as originally planned, and major portions of each second-stage form were reused.

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. To contribute useful information about children's skill levels, test items selected for the final forms should ideally have high *r*-biserials (.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 prior-round ECLS-K:2011 national assessments or the ECLS-K national assessments, or links to a selected reading passage.

4.2 Reading

Overall, the reading field test items and the items from the ECLS-K:2011 and the ECLS-K national administrations of 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 .3. The items showed the expected trends in response selection in that 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 for most of the items. In selecting items for the national first- and second-grade assessments, items with low *r*-biserials, poor fit, or low discrimination were avoided.

The EBRS items included in the kindergarten assessments were re-evaluated for inclusion in the first-grade assessments as part of the full reading assessment item evaluation. The majority of the EBRS items performed well in the field test and in the ECLS-K:2011 and the ECLS-K national assessments. Results of the item evaluation indicated it would be appropriate to administer the same EBRS assessment in first grade that was administered in kindergarten, with no changes. Item difficulties for the set of EBRS were well below the levels estimated for second grade; therefore, the EBRS item set was not administered after first grade.

4.2.1 Estimated Ability Levels for the ECLS-K:2011 National Samples and Target Ranges for Item Difficulties

Table 4-1 provides the estimated first-grade means and standard deviations of ability level (theta), all calibrated on the same scale, for the different samples in the pooled analysis described above. This information was used to calculate the full range of ability levels (and, therefore item difficulties) that needed to be covered by the first-grade reading assessment in the ECLS-K:2011. Table 4-2 provides the same information for the second-grade reading assessment. Note that the values in table 4-1 are *not* comparable to those in table 4-2; the theta estimates in these tables are not on a common scale because the values were estimated using the different sets of pooled data described above in section 4.1.2.

Tables 4-3 and 4-4 show the estimated ability ranges for the overall assessment as well as for the low-, middle-, and high-ability level groups for the fall and spring of first grade and for the fall and spring of second grade. Tables 4-3 and 4-4 also show the number of items selected for each 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 to avoid floor and ceiling effects and to provide additional overlap between forms to support development of a common score scale. Note that the ability ranges in tables 4-3 are not directly comparable to the ranges in table 4-4 since the thetas were estimated using the different sets of pooled data described above.

Table 4-1. Means and standard deviations of estimated reading ability level (theta) for children in first grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Fall first grade – ECLS-K:2011 field test	0.34	0.53
Fall first grade – ECLS-K national data collection	-0.13	0.59
Spring first grade – ECLS-K:2011 field test[1]	0.96	0.53
Spring first grade – ECLS-K national data collection	0.51	0.52

[1] Without data from the field test for spring first grade, an estimate of the spring first-grade ECLS-K:2011 field test mean was calculated by assuming the growth (in standard deviation units) from fall first grade to spring first grade in the ECLS-K:2011 field test would be approximately the same as the growth from fall first grade to spring first grade in the ECLS-K national data collection. The standard deviation from the fall first-grade ECLS-K:2011 field test was assumed to be the same at spring first grade.

NOTE: The values in this table are not comparable to the values in table 4-2 because the thetas are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2009 field test, and Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1999 and spring 2000.

Table 4-2. Means and standard deviations of estimated reading ability level (theta) for children in second grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Fall second grade – ECLS-K:2011 field test	0.95	0.47
Fall second grade – ECLS-K national data collection[1]	0.49	0.55
Spring second grade – ECLS-K:2011 field test[2]	1.30	0.47
Spring second grade – ECLS-K national data collection[3]	0.90	0.55

[1] Without data from the ECLS-K national data collection for second grade, an estimate of the fall second-grade ECLS-K national mean was assumed to be the same as the spring first-grade ECLS-K national mean. The standard deviation from the fall second-grade ECLS-K national mean was also assumed to be the same as the standard deviation at spring first grade.

[2] Without data from the field test, the spring second-grade ECLS-K:2011 field test mean was assumed to be the same as the fall third-grade field test mean theta estimate. The standard deviation for the spring second-grade ECLS-K:2011 field test was assumed to be the same as the standard deviation at fall second grade.

[3] Without data from the ECLS-K national data collection for second grade, an estimate of the spring second-grade ECLS-K national mean was calculated by assuming the growth (in standard deviation units) from fall second grade to spring second grade in the ECLS-K:2011 field test would be approximately the same as the growth from fall second grade to spring second grade in the ECLS-K national data collection. The standard deviation from the fall second-grade ECLS-K national estimate was assumed to be the same at spring second grade.

NOTE: The values in this table are not comparable to the values in table 4-1 because the thetas are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2009 field test, and Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2000 and spring 2002.

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

	Fall first grade			Spring first grade		
Item	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2*SD*)	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2*SD*)
Estimated ability range	-1.30 to +0.34	-0.71 to +0.88	-0.13 to +1.41	-0.52 to +0.96	-0.01 to +1.49	+0.51 to +2.02
Number of items with difficulties in estimated peak ability range:						
Routing	17	24	21	23	18	7
Low form	9	7	4	6	3	0
Middle form	4	14	19	14	20	14
High form	0	1	10	4	10	17

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in table 4-4 because the abilities and difficulties are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011 and spring 2012.

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

	Fall second grade			Spring second grade		
Item	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2*SD*)	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2*SD*)
Estimated ability range	-0.61 to +0.95	-0.06 to +1.42	+0.49 to +1.89	-0.20 to +1.30	+0.35 to +1.77	+0.90 to +2.24
Number of items with difficulties in estimated peak ability range:						
Routing	20	23	16	19	17	9
Low form	14	19	14	19	19	7
Middle form	13	19	19	20	21	11
High form	4	10	16	9	15	15

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in table 4-3 because the abilities and difficulties are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012 and spring 2013.

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. Also, only a limited number of passages could be included in any assessment form, because the time for 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.

Another component of the assessment design included adherence to the framework specifications described in chapter 2. 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, the passages also had associated items in the **locate/recall** and **vocabulary** categories. 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 4-5 provides information about how the final reading assessments developed for the national first-grade and second-grade rounds of data collection compare to the framework specifications in terms of the distribution of items by content category. The kindergarten information is included for comparison. 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.

Table 4-5. Framework targets and items by content area for the national kindergarten, first-grade, and second-grade reading assessments: ECLS-K:2011

	Kindergarten			First grade			Second grade		
Content area	Targeted percent of items	Actual number	Actual percent	Targeted percent of items	Actual number	Actual percent	Targeted percent of items	Actual number	Actual percent
Total	100	83	100	100	100	100	100	73	100
Basic reading skills	50	53	64	40	51	51	20	16	22
Vocabulary	15	11	13	15	12	12	10	10	14
Comprehension									
Locate/recall	20	14	17	20	19	19	30	22	30
Integrate/ interpret	10	3	4	20	13	13	30	19	26
Critique/ evaluate	5	2	2	5	5	5	10	6	8

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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

The passage sets were selected to maximize the number of **integrate/interpret** and **critique/evaluate** items of appropriate difficulty for the each of the assessments. However, as can be seen in the information presented in table 4-5, even with this maximization, the percentage of items in each of these categories sometimes fell short of the targets. The available item pools did not include enough items in these categories that performed well at the expected grade levels. The percentages of items in the **locate/recall** and **vocabulary** categories fell at or close to the targeted percentages at each grade level, with the percentage of **vocabulary** items in second grade higher than targeted due to the inclusion of items in passage sets. The percentage of items from the **basic skills** category was higher than targeted at all grade levels, even more so at kindergarten and first grade when the EBRS items are included in the counts, resulting in a deviation from the targets in the assessment framework specifications.

One way the reading assessments differ from those in the mathematics and science domains is in how items associated with passage sets, or of the same item type, were ultimately calibrated and scored. Unlike the mathematics and science assessments, which included only dichotomous items, sets of reading items were treated as single polytomous items, with a range of values from 0 to the number of items in the set. For example, for a set of four items associated with a passage, a single item score was developed that represented the number of items correctly answered by the child in that particular set, which in this case, would result in an item score ranging from 0–4. Thus, polytomous items carried with them score points greater than 1, as with dichotomous items. In the kindergarten reading assessment, 13 item sets were defined, including, for example, a set of items assessing knowledge of letter sounds and a set of items associated with a reading passage. In the first grade reading assessment, 16 item sets were defined, and 8 item sets were defined in the second-grade assessment, Four items sets in the kindergarten assessment, half of the item sets in the first-grade assessment, and all of the items sets in the second-grade assessment were associated with reading passages. The number of dichotomous and polytomous items, and for the latter, the number of score points for the kindergarten, first- and second-grade reading assessments are summarized in table 4-6.

Table 4-6. Dichotomous and polytomous items and score values for the national kindergarten, first-grade, and second-grade reading assessments: ECLS-K:2011

Number of items	Kindergarten	First grade	Second grade
Total	56	62	42
Dichotomous	43	46	33
Polytomous			
2-point	5	5	0
3-point	3	2	0
4-point	4	7	6
5-point	1	2	2
6-point	0	0	1

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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

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 .3, and 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 in the kindergarten psychometric report, 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. The SERS assessment administered in first grade was the same one administered in kindergarten, with no changes. Since some higher difficulty items with *b* parameter estimates more than two standard deviations higher than the estimated mean were included to avoid ceiling effects in the kindergarten assessment, it was still appropriate for use in first grade. Also, unlike the assessments in English, it was not expected that the average SERS test taker would acquire a higher level of Spanish language knowledge and skills between kindergarten and first grade, given that the primary instruction received in the classroom was expected to be in English.

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 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 kindergarten reading assessment. The resulting percentage of vocabulary items was higher than targeted, while the percentage of basic skills items was lower than targeted, a result of the number of items available in the pool that were appropriate to translate into Spanish.

Table 4-7. Framework targets and items by content area for the national 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	83	24	77
Vocabulary	17	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), fall 2010 and fall 2011.

4.3 Mathematics

Overall, the mathematics field test items and the items from the ECLS-K:2011 and the ECLS-K national administrations of the mathematics assessment performed well. As in the reading domain, the item analysis showed that the majority of items had *r*-biserials that were well above the desired value of .3, and the IRT plots showed good fit of item data with the estimated parameters for most of the items. In selecting items for the national first- and second-grade assessments, items with low *r*-biserials, poor fit, or low discrimination were avoided.

4.3.1 Estimated Ability Levels for the ECLS-K:2011 National Samples and Target Ranges for Item Difficulties

Table 4-8 provides the first-grade means and standard deviations of the estimated ability level (theta), all calibrated on the same scale, for the different samples in the pooled analysis described above. This information was used to calculate the full range of ability levels (and, therefore item difficulties) that needed to be covered by the first-grade mathematics assessment in the ECLS-K:2011. Table 4-9 provides the same information for the second-grade mathematics assessment. The values in table 4-8 are *not* comparable to those in table 4-9; the theta estimates in these tables are not on a common scale because the values were estimated using the different sets of pooled data described in section 4.1.2.

Tables 4-10 and 4-11 show the estimated ability ranges for the entire assessment as well as for the low-, middle-, and high-ability level groups for the fall and spring of first grade and the fall and spring of second grade, respectively. Tables 4-10 and 4-11 also show the number of items selected for each 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. Note that the ability ranges in tables 4-10 and 4-11 are not directly comparable to one another since the thetas for the first-grade assessment and for the second-grade assessment were estimated using the different sets of pooled data described above.

Table 4-8. Means and standard deviations of estimated mathematics ability level (theta) for children in first grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Fall first grade – ECLS-K:2011 field test	0.27	0.49
Fall first grade – ECLS-K national data collection	-0.10	0.60
Spring first grade – ECLS-K:2011 field test[1]	0.78	0.49
Spring first grade – ECLS-K national data collection	0.49	0.55

[1] Without data from the field test for spring first grade, an estimate of the spring first-grade ECLS-K:2011 field test mean was calculated by assuming the growth (in standard deviation units) from fall first grade to spring first grade in the ECLS-K:2011 field test would be approximately the same as the growth from fall first grade to spring first grade in the ECLS-K national data collection. The standard deviation from the fall first-grade ECLS-K:2011 field test was assumed to be the same at spring first grade.

NOTE: The values in this table are not comparable to the values in table 4-8 because the thetas are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2009 field test, and Kindergarten Class of 1998–99 (ECLS-K), fall 1999 and spring 2000.

Table 4-9. Means and standard deviations of estimated mathematics ability level (theta) for children in second grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Fall second grade – ECLS-K:2011 field test	0.72	0.45
Fall second grade – ECLS-K national data collection[1]	0.38	0.57
Spring second grade – ECLS-K:2011 field test[2]	1.17	0.46
Spring second grade – ECLS-K national data collection[3]	0.95	0.57

[1] Without data from the ECLS-K national data collection for second grade, an estimate of the fall second-grade ECLS-K national mean was assumed to be the same as the spring first-grade ECLS-K national mean. The standard deviation from the fall second-grade ECLS-K national mean was also assumed to be the same as the standard deviation at spring first grade.

[2] Without data from the field test, the spring second-grade ECLS-K:2011 field test mean was assumed to be the same as the fall third-grade field test mean theta estimate. The standard deviation for the spring second-grade ECLS-K:2011 field test was assumed to be the same as the standard deviation at fall second grade.

[3] Without data from the ECLS-K national data collection for second grade, an estimate of the spring second-grade ECLS-K national mean was calculated by assuming the growth (in standard deviation units) from fall second grade to spring second grade in the ECLS-K:2011 field test would be approximately the same as the growth from fall second grade to spring second grade in the ECLS-K national data collection. The standard deviation from the fall second-grade ECLS-K national estimate was assumed to be the same at spring second grade.

NOTE: The values in this table are not comparable to the values in table 4-7 because the thetas are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2009 field test, and Kindergarten Class of 1998–99 (ECLS-K), spring 2000 and spring 2002.

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

	Fall first grade			Spring first grade		
Item	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2SD)	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2*SD*)
Estimated ability range	-1.30 to +0.27	-0.70 to +0.77	-0.10 to +1.26	-0.60 to +0.78	-0.05 to +1.27	+0.49 to +1.77
Number of items with difficulties in estimated peak ability range:						
Routing	11	10	6	6	4	4
Low form	18	15	4	14	3	2
Middle form	10	17	24	17	23	18
High form	0	3	15	3	16	26

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in table 4-11 because the abilities and difficulties are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011 and spring 2012.

Table 4-11. Peak difficulty ranges for the national second-grade mathematics assessment, routing plus second stage: ECLS-K:2011

	Fall second grade			Spring second grade		
Item	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2SD)	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2*SD*)
Estimated ability range	-0.76 to +0.72	-0.19 to +1.17	+0.38 to +1.62	-0.19 to +1.17	+0.38 to +1.63	+0.95 to +2.09
Number of items with difficulties in estimated peak ability range:						
Routing	11	10	7	10	7	4
Low form	19	20	17	20	17	4
Middle form	4	17	25	17	25	21
High form	0	9	15	9	15	20

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in table 4-10 because the abilities and difficulties are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012 and spring 2013.

As discussed in chapter 2, adherence to the framework category targets in mathematics was required during assessment design. Table 4-12 provides information about how the final mathematics assessments developed for the national first- and second-grade rounds of data collection compared to the framework specifications in terms of the distribution of items by content category. The kindergarten item distribution is included for comparison.

The actual percentages of items matched or were very close to the targeted percentages in all categories at all grade levels. Any shortfalls were due to the lack of items in the item pool that fell within the difficulty range suitable for the grade level targeted and had good psychometric characteristics. Conversely, some categories had more items than targeted so that the distribution of item difficulties across the expected range within a second-stage form would ensure accurate measurement across the ability distribution.

Table 4-12. Framework targets and items by content area for the national kindergarten, first-grade, and second-grade mathematics assessments: ECLS-K:2011

	Kindergarten			First grade			Second grade		
Content area	Targeted percent of items	Actual number	Actual percent	Targeted percent of items	Actual number	Actual percent	Targeted percent of items	Actual number	Actual percent
Total	100	75	100	100	85	100	100	75	100
Number properties and operations	75	57	76	75	63	74	75	55	73
Measurement	5	2	3	5	4	5	5	4	5
Geometry	3	2	3	3	2	2	3	2	3
Data analysis and probability	8	6	8	8	6	7	8	6	8
Algebra	9	8	11	9	10	12	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), fall 2010, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

4.4 Science

Overall, the science items administered to first- and second-graders in the 2009 field test, the items from the ECLS-K:2011 national administrations of the science assessment in kindergarten, and the science items administered in the K-1 general knowledge assessment and the third-grade science assessment in the ECLS-K national data collections performed well. As in the reading domain, the item analysis showed that the majority of items had *r*-biserials that were well above the desired value of .3, and the IRT plots showed good fit of item data with the estimated parameters for most of the items. In selecting items for the national first- and second-grade assessments, items with low *r*-biserials, poor fit, or low discrimination were avoided.

4.4.1 Estimated Ability Levels for the ECLS-K:2011 National Samples and Target Ranges for Item Difficulties

Table 4-13 provides the first-grade means and standard deviations of the estimated ability level (theta), all calibrated on the same scale, for the different samples in the pooled analysis described above. This information was used to calculate the full range of ability levels (and, therefore item difficulties) that needed to be covered by the first-grade science assessment in the ECLS-K:2011. Table 4-14 provides the same information for the second-grade science assessment. The values in table 4-13 are *not* comparable to those in table 4-14; the theta estimates in these tables are not on a common scale because the values were estimated using the different sets of pooled data described in section 4.1.2.

Tables 4-15 and 4-16 show the estimated ability ranges for the overall assessment as well as for the low-, middle-, and high-ability level groups for the fall and spring of first grade and for the fall and spring of second grade, respectively. Tables 4-15 and 4-16 also show the number of items selected for each 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 reading and mathematics assessments, items with difficulty parameter estimates below the anticipated lowest theta and above the anticipated highest theta were included to avoid floor and ceiling effects. Note that the ability ranges in tables 4-15 and 4-16 are not directly comparable to one another since the thetas for the first-grade assessment and the second-grade assessment were estimated using the different sets of pooled data described above.

Table 4-13. Means and standard deviations of estimated science ability level (theta) for children in first grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Fall first grade – ECLS-K:2011 field test	0.28	0.72
Fall first grade – ECLS-K national data collection	0.23	0.83
Spring first grade – ECLS-K:2011 field test[1]	0.58	0.72
Spring first grade – ECLS-K national data collection	0.57	0.80

[1] Without data from the field test for spring first grade, an estimate of the spring first-grade ECLS-K:2011 field test mean was calculated by assuming the growth (in standard deviation units) from fall first grade to spring first grade in the ECLS-K:2011 field test would be approximately the same as the growth from fall first grade to spring first grade in the ECLS-K national data collection. The standard deviation from the fall first-grade ECLS-K:2011 field test was assumed to be the same at spring first grade.

NOTE: The values in this table are not comparable to the values in table 4-14 because the thetas are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2009 field test, and Kindergarten Class of 1989–99 (ECLS-K), fall 1999 and spring 2000.

Table 4-14. Means and standard deviations of estimated science ability level (theta) for children in second grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Fall second grade – ECLS-K:2011 field test	0.51	0.74
Fall second grade – ECLS-K national data collection[1]	0.10	0.78
Spring second grade – ECLS-K:2011 field test[2]	1.00	0.74
Spring second grade – ECLS-K national data collection[3]	0.61	0.78

[1] Without data from the ECLS-K national data collection for second grade, an estimate of the fall second-grade ECLS-K national mean was assumed to be the same as the spring first-grade ECLS-K national mean. The standard deviation from the fall second-grade ECLS-K national mean was also assumed to be the same as the standard deviation at spring first grade.

[2] Without data from the field test, the spring second-grade ECLS-K:2011 field test mean was assumed to be the same as the fall third-grade field test mean theta estimate. The standard deviation for the spring second-grade ECLS-K:2011 field test was assumed to be the same as the standard deviation at fall second grade.

[3] Without data from the ECLS-K national data collection for second grade, an estimate of the spring second-grade ECLS-K national mean was calculated by assuming the growth (in standard deviation units) from fall second grade to spring second grade in the ECLS-K:2011 field test would be approximately the same as the growth from fall second grade to spring second grade in the ECLS-K national data collection. The standard deviation from the fall-second grade ECLS-K national estimate was assumed to be the same at spring second grade.

NOTE: The values in this table are not comparable to the values in table 4-13 because the thetas are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2009 field test, and Kindergarten Class of 1998–99 (ECLS-K), spring 2000 and spring 2002.

Table 4-15. Peak difficulty ranges for the national first-grade science assessment, routing plus second stage: ECLS-K:2011

	Fall first grade			Spring first grade		
Item	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2SD)	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2*SD*)
Estimated ability range	-1.43 to +0.28	-0.60 to +1.05	+0.23 to +1.88	-1.04 to +0.58	-0.24 to +1.37	+0.57 to +2.17
Number of items with difficulties in estimated peak ability range:						
Routing	6	9	9	7	9	8
Low form	9	9	5	8	7	4
Middle form	8	9	8	8	8	5
High form	4	8	9	5	7	8

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in table 4-16 because the abilities and difficulties are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011 and spring 2012.

Table 4-16. Peak difficulty ranges for the national second-grade science assessment, routing plus second stage: ECLS-K:2011

	Fall second grade			Spring second grade		
Item	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2SD)	Estimated low-level abilities (-2*SD* to mean)	Estimated mid-level abilities (-1*SD* to +1*SD*)	Estimated high-level abilities (mean to +2*SD*)
Estimated ability range	-1.46 to +0.51	-0.68 to +1.25	+0.10 to +1.99	-0.95 to +1.00	-0.17 to +1.74	+0.61 to +2.48
Number of items with difficulties in estimated peak ability range:						
Routing	10	9	7	10	10	6
Low form	10	9	7	10	9	5
Middle form	5	9	9	7	8	10
High form	0	5	7	3	6	12

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in table 4-15 because the abilities and difficulties are not on the same scale. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 20 12 and spring 2013.

Table 4-17 provides information about how the final science assessments developed for the national first- and second-grade rounds of data collection compared to the framework specifications in terms of the distribution of items by content category. The kindergarten item distribution is included for comparison.

The actual percentages of items match the targets exactly in all categories in kindergarten. In the first and second grades, shortfalls were due to the lack of items in the item pool that fell within the difficulty range suitable for the grade level targeted and had good psychometric characteristics. Conversely, some categories had more items than targeted so that the distribution of item difficulties within a second-stage form and across the expected range would ensure accurate measurement across the ability distribution.

Table 4-17. Framework targets and items by content area for the national kindergarten, first-grade, and second-grade science assessments: ECLS-K:2011

	Kindergarten			First grade			Second grade		
Content area	Targeted percent of items	Actual number	Actual percent	Targeted percent of items	Actual number	Actual percent	Targeted percent of items	Actual number	Actual percent
Total	100	20	100	100	43	100	100	48	100
Scientific inquiry	25	5	25	25	12	28	25	12	25
Physical science	25	5	25	25	12	28	25	13	27
Life science	25	5	25	25	10	23	25	11	23
Earth and space science	25	5	25	25	9	21	25	12	25

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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

4.5 Cut Scores Used for Routing

As noted earlier, the first- and second-grade assessments were developed at separate points in time. For each assessment, once the items were selected and allocated to the routing and low-, middle-, and high-level second-stage tests for the national assessments, simulations of performance on the routing and second-stage tests were run in order to calculate the cut scores for the routing test that would determine which second-stage form children would be administered. To conduct the simulations used to determine the cut scores for each grade for each domain, 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 ability levels at each grade level 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 .9, 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 for first grade included simulations on data from four samples: (1) the fall of first grade from the ECLS-K:2011 field test, (2) the fall of first grade from the ECLS-K national data collection, (3) the spring of first grade from the ECLS-K:2011 field test (interpolated from the fall of kindergarten and the fall of first-grade data), and (4) the spring of first grade from the ECLS-K national data collection. The cut score simulations for second grade also included data from four samples: (1) the fall of second grade from the ECLS-K:2011 field test, (2) the fall of second grade from the ECLS-K national data collection (interpolated from the spring of first-grade and the spring of third-grade data), (3) the spring of second grade from the ECLS-K:2011 field test (interpolated from the fall of first-grade and the fall of third-grade data), and (4) the spring of second grade from the ECLS-K national data collection (interpolated from the spring of first-grade and the spring of third-grade data).

The estimated numbers of floor and ceiling occurrences also were 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 router and low forms was determined. If this number was less than 3 percent of the sample then 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 router and high forms was less than 3 percent, that would have been taken as evidence of a negligible ceiling effect.

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 router 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 entirely 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.

4.5.1 Reading

For the reading assessments, cut scores were analyzed for the routing form; therefore, four simulations were performed for reading, one for each routing form within each sample for each grade. The reading simulations at each grade level showed no evidence of a significant floor or ceiling effect using any of the samples. For the first-grade reading assessment, the analysis of optimal cut scores indicated that children should be routed directly to the low second-stage form if they had a router score of 12 or lower (including 0). Children who scored between 13 and 23 items correct on the router would proceed with the middle form, while those with scores of 24 or higher would proceed to the high form.

For the second-grade reading assessment, the analysis of optimal cut scores indicated that children with a router score of 17 or lower should be directed to the low form, while children with router scores of 18 to 24 and 25 to 29 should be directed to the middle and high forms, respectively.

4.5.2 Mathematics

For the first-grade mathematics assessment, the analysis of optimal cut scores indicated that children should be routed directly to the low second-stage form if they had a router score of 7 or lower.

Children who scored between 8 and 12 items correct on the router would proceed with the middle form, while those with scores of 13 or higher would proceed to the high form.

For the second-grade mathematics assessment, the analysis of optimal cut scores indicated that children with a router score of 12 or lower should be directed to the low form, while children with router scores of 13 to 16 and 17 to 20 should be directed to the middle and high forms, respectively.

4.5.3 Science

For the first-grade science assessment, the analysis of optimal cut scores indicated that children should be routed directly to the low second-stage form if they had a router score of 4 or lower. Children who scored between 5 and 9 items correct on the router would proceed with the middle form, while those with scores of 10 or higher would proceed to the high form.

For the second-grade science assessment, the analysis of optimal cut scores indicated that children with a router score of 10 or lower should be directed to the low form, while children with router scores of 11 to 14 and 15 to19 should be directed to the middle and high forms, respectively.

4.6 Discontinue Rules

Once the assessment forms were finalized and the cut scores were determined, discontinue rules were implemented in some of the assessments to further guard against children being administered items that were much too difficult for them, given their performance on items earlier in the assessment. Also, it was noted during field observations in the data collections that children tended to take more time on the more difficult items. The discontinue rules served to limit both the frustration that could stem from being given items that were too hard and the length of time children would spend trying to provide an answer for items they were unlikely to answer correctly. For example, in the reading assessment, children who answer the sight word items incorrectly most likely would not be able to answer the sentence reading items correctly. Therefore, children who were not able to read words would not be administered the sentence reading items. Or if a child responded incorrectly to many or all of the items associated with an easy passage set, then the subsequent more difficult passage sets, which would take them much longer to read, would not be administered. In the mathematics assessment, for example, children who answered the addition and subtraction items incorrectly most likely would not be able to answer the multiplication and division items correctly.

Other discontinue rules in the mathematics assessment were empirically determined based on an analysis of item difficulties in sections of the assessment along with the estimated ability levels of children in the fall and spring administrations. For example, on the second-stage high form in first grade (which was the same as the second-stage middle form in second grade), the difficulties of the first 20 items administered ranged from approximately 1.0 to 2.5 standard deviations above the estimated mean ability level in the fall administration, and from approximately the mean to 1.5 standard deviations above the estimated mean ability level in the spring administration. The difficulties of the remaining items ranged from 2.7 to 3.6 standard deviations above the estimated fall mean ability level and 1.6 to 2.6 standard deviations above the estimated spring mean ability level. Based on these estimates and guided by professional judgments about burden and the potential for a child to become too frustrated, discontinue rules were defined by the likelihood of a child responding correctly to any of the subsequent, more difficult items on a form, considering the child responded incorrectly to a majority or all of the easier items in the defined section on that same form.

4.6.1 Reading

In the first-grade reading assessment, four discontinue rules were implemented:

1. On the routing form, if the first four sight word and the first four sentence reading items were answered incorrectly, the router was ended and the child was directed to a second-stage form.
2. On the routing form, if all three of the items associated with the first passage were answered incorrectly, the router was ended and the child was directed to a second-stage form.
3. On the second-stage middle form, if all four items associated with the first passage were answered incorrectly, the reading assessment was discontinued.
4. On the second-stage middle form, if all four items associated with the second passage were answered incorrectly, the reading assessment was discontinued.

In the second-grade reading assessment, six discontinue rules were implemented:

1. On the routing form, if at least two of the three items associated with the first passage were answered incorrectly, the router was ended and the child was directed to a second-stage form.

2. On the routing form, if at least two of the four items associated with the second passage were answered incorrectly, the router was ended and the child was directed to a second-stage form.

3. On the second-stage low form, if all four items associated with the first passage were answered incorrectly, the reading assessment was discontinued.

4. On the second-stage low form, if all four items associated with the second passage were answered incorrectly, the reading assessment was discontinued.

5. On the second-stage high form, if all nine items associated with the first and second passages (four items for the first passage and five items for the second passage) were answered incorrectly, the reading assessment was discontinued.

6. On the second-stage high form, if at least two of the four items associated with the third passage were answered incorrectly, the reading assessment was discontinued.

4.6.2 Mathematics

In the first-grade mathematics assessment, two discontinue rules were implemented:

1. On the second-stage high form, if 13 or more of the first 20 items were answered incorrectly, the mathematics assessment was discontinued.

2. On the second-stage high form, if the 11th through 20th items were all answered incorrectly, the mathematics assessment was discontinued.

In the second-grade mathematics assessment, three discontinue rules were implemented:

1. On the second-stage low form, if at least four of the first eight items were answered incorrectly, the mathematics assessment was discontinued.

2. On the second-stage middle form, if at least 13 of the first 20 items were answered incorrectly, the mathematics assessment was discontinued.

3. On the second-stage middle form, if the 11th through 20th items were all answered incorrectly, the mathematics assessment was discontinued.

4.6.3 Science

Discontinue rules were not included in the science assessment for either the first or second grade. This was because the knowledge and skills measured by items that appeared later in the assessments did not necessarily build on the knowledge and skills assessed by items earlier in the assessment. For example, a child who knows little about the planets and space but a lot about plants and animals might not be able to answer an easier item in the content area of Earth science but have no difficulties answering questions about life science, including those that are more difficult. Additionally, most items in the science assessments were relatively quick and easy to administer, so there was less concern about children spending too much time thinking about the answer to any particular item.

This page intentionally left blank.

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, fall 2011 and spring 2012 first-grade, and fall 2012 and spring 2013 second-grade rounds of the ECLS-K:2011. Although the focus of this report is predominantly on the psychometric results for the first- and second-grade data collections, results from the kindergarten data collection rounds are also included in this chapter to provide the reader with the complete longitudinal analysis results. Additionally, results from the kindergarten round are included in this chapter due to the change in scoring the reading assessment (from dichotomous to polytomous calibration, see section 5.2.4.1 for more information). 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 batteries, which is important information for understanding the specific scores that have been developed. Section 5.2 includes the approach to scoring the assessment and the 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 the selection and use of the scores in section 5.8, with section 5.8.1 focusing on choosing the best scores for certain types of analyses and section 5.8.2 noting some important considerations when using assessment scores to measure gain.

5.1 Routing of Children Through the National Assessments

The full direct assessment batteries included assessments in reading, mathematics, science, and executive function,[1] as well as measurements of height and weight.

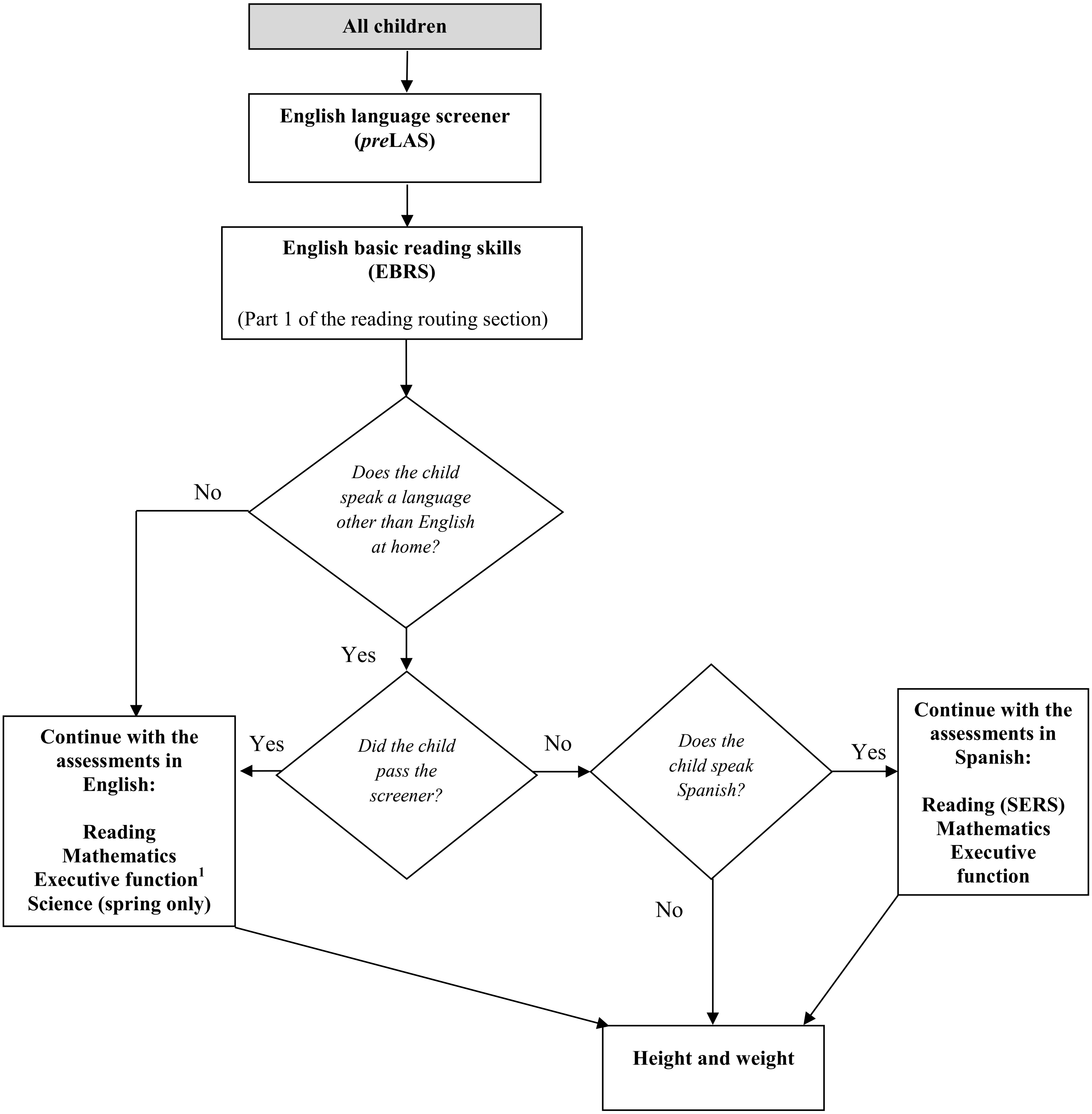
[1] Executive functions are interdependent processes that work together to regulate and orchestrate cognition, emotion, and behavior and that help a student learn in the classroom (e.g., Diamond 2013). Two measures of executive function were administered in fall and spring kindergarten, fall and spring first-grade, and fall and spring second-grade. The *Dimensional Change Card Sort* (DCCS), which measures cognitive flexibility was administered first, followed by Numbers Reversed, which measures working memory. Executive function measures are discussed in more detail in chapter 6.

5.1.1 Kindergarten Rounds

As illustrated in exhibit 5-1, prior to being administered the reading assessment in the kindergarten rounds of data collection, 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 one 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 required children to give the name of a pictured object, which tested children's expressive vocabulary. Performance on the ECLS-K:2011 language screener determined which components of the assessment a child received.

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 basic sight word 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 who did not respond correctly to at least 10 of the 20 EBRS items were routed directly to the low-level second-stage test. 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 (*Dimensional Change Card Sort [DCCS]* and Numbers Reversed) 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.

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


All children
English language screener (*pre*LAS)
English basic reading skills (EBRS)
(Part 1 of the reading routing section)
Does the child speak a language other than English at home?
No
Yes
Did the child pass the screener?
Yes
No
Does the child speak Spanish?
Yes
No
Continue with the assessments in English:
Reading
Mathematics
Executive function[1]
Science (spring only)
Continue with the assessments in Spanish:
Reading (SERS)
Mathematics
Executive function
Height and weight

[1] Two measures of executive function were measured. The *DCCS* was administered first, followed by Numbers Reversed.

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.

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, the children's height and weight were measured. Children whose primary home language was not Spanish were routed out of the rest of the cognitive assessments, and their height and weight were measured.

5.1.2 First-Grade Rounds

In the first-grade data collections, children were assessed in reading, mathematics, and science in both the fall and the spring. All children received the assessments designed for the first-grade collections, regardless of their actual grade level. In both the fall and the spring, students' executive function skills were assessed with the same measures fielded in kindergarten, the *DCCS* and Numbers Reversed. Finally, children's height and weight were measured again in both fall and spring.

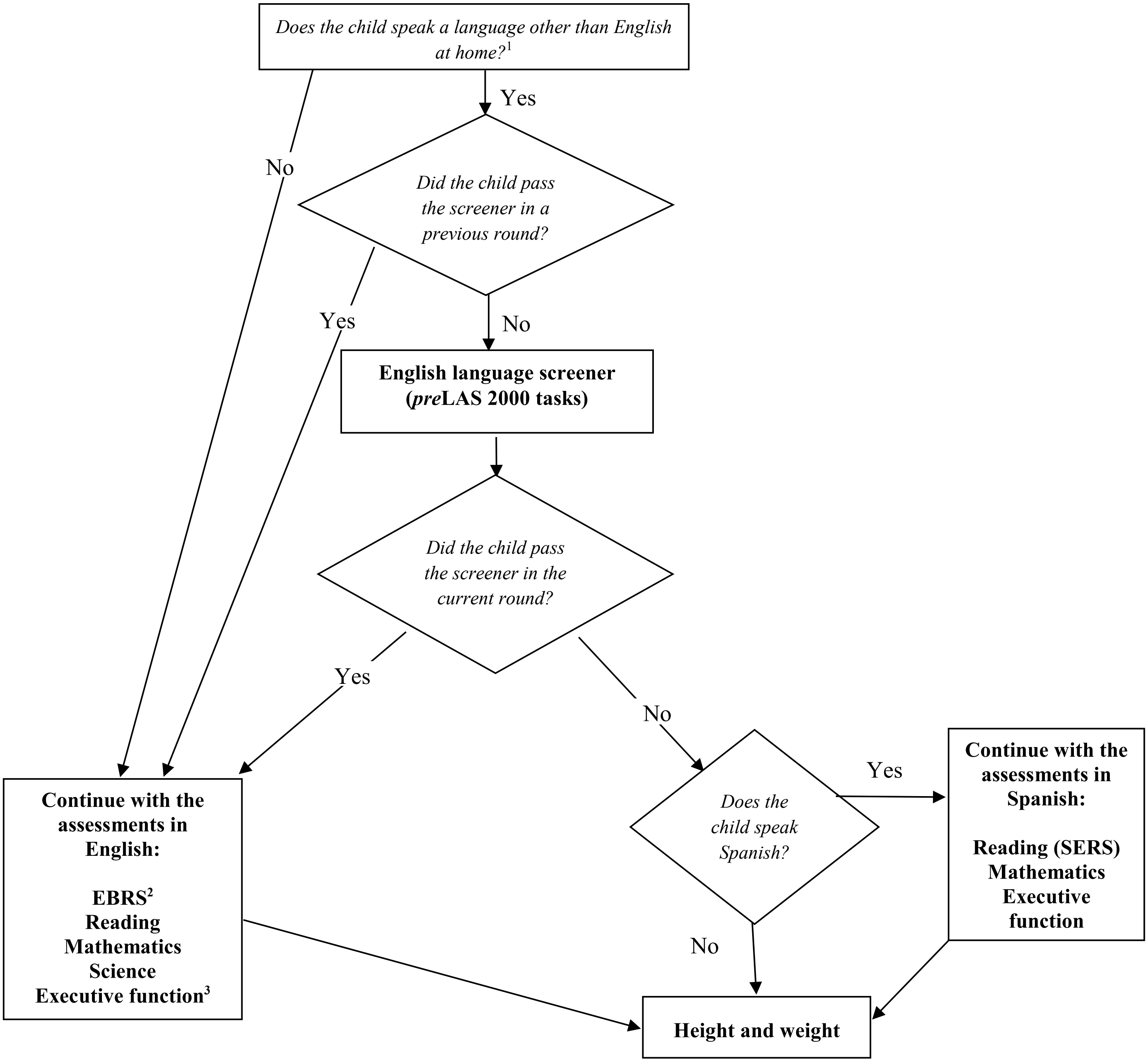
The components of the ECLS-K:2011 assessments administered to children who spoke a language other than English at home (as determined in the kindergarten year using information collected from school records or school staff) depended on the children's performance on the same language screener used in the kindergarten rounds. In the fall and spring kindergarten rounds, all children were administered the language screener as the first component of the direct cognitive assessment, regardless of their home language. In contrast to the procedures used in kindergarten, the screener was not administered to all children in the first-grade collections. The two *pre*LAS 2000 tasks were administered only to children who spoke a language other than English at home and had not passed the screener in the most recent round in which they were assessed. For example, children who spoke a language other than English at home who were assessed most recently in the spring of kindergarten and did not pass the *pre*LAS screener at that time were administered the screener the next time they were assessed. Such children who were part of the fall first-grade subsample were administered the *pre*LAS screener in the fall of first grade. If they did not pass the screener in the fall, it was administered to them again in the spring. Children who were not part of the fall subsample, spoke a language other than English at home, and did not achieve at least a minimum score on the screener in the spring of kindergarten were administered the *pre*LAS screener in the spring of first grade. Children who were not administered the language screener either because they did not speak a language other than English at home or because they passed the screener in a previous round were asked only two of the *pre*LAS "Art Show" items as a warm-up; they were not administered any of the other *pre*LAS items.

In the first-grade data collections, all children who were routed to the English version of the assessment were administered a 30-item reading routing test. Depending on the number of correct responses a child provided to items on the reading routing test, he or she was routed to one of three second-stage reading tests. Those children whose scores routed them to the low or middle second-stage tests in reading first received 18 items that contribute to the calculation of the EBRS score.[2] After administration of these 18 items, children proceeded into the low or middle second-stage test. Children who were routed to the high second-stage test based on their scores on the 30-item router were not administered the 18 items that contribute to the EBRS because these items were considered too easy for their demonstrated ability level. Once the reading assessments were complete, the mathematics, science, and executive function measures were administered in English, followed by measurements of height and weight.

Children who were administered the two *pre*LAS 2000 tasks in first grade and did not achieve at least the minimum score on the language screener were administered the 18 EBRS items after the screener. Once the EBRS items were administered, the cognitive assessments in English ended for these children. Spanish-speaking children who did not achieve at least the minimum score on the screener were then administered the SERS and the mathematics and executive function assessments that had been translated into Spanish. Children whose home language was one other than English or Spanish and did not achieve at least the minimum score on the screener were not administered any of the remaining cognitive assessments, although all children had their height and weight measured. Exhibit 5-2 illustrates how the first-grade assessments taken by children depended on their home language and on their performance on the language screener.

[2] The EBRS provides information on children's performance on these 18 items plus the 2 items from the *pre*LAS "Art Show" task that were administered to all children at the beginning of the assessment.

Exhibit 5-2. Routing path for the direct child assessment in the ECLS-K:2011 first-grade year


Does the child speak a language other than English at home?[1]
Yes
No
Did the child pass the screener in a previous round?
Yes
No
English language screener (*pre*LAS 2000 tasks)
Did the child pass the screener in the current round?
Yes
No
Does the child speak Spanish?
Yes
No
Continue with the assessments in English:
EBRS[2]
Reading
Mathematics
Science
Executive function[3]
Continue with the assessments in Spanish:
Reading (SERS)
Mathematics
Executive function
Height and weight

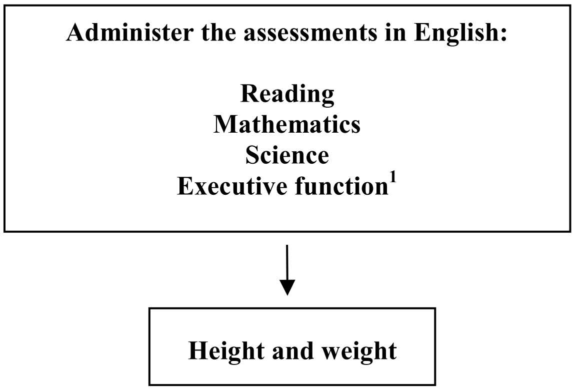
[1] Home language designation was identified in the kindergarten rounds of data collection.
[2] The EBRS was administered in the English reading battery only to children who were routed to the low and middle second-stage reading forms.
[3] Two measures of executive function were administered. The DCCS was administered first, followed by Numbers Reversed.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2011 and spring 2012.

5.1.3 Second-Grade Rounds

In the second-grade data collections, children were again assessed in reading, mathematics, and science in both the fall and the spring. All children received the assessments designed for the second-grade collections, regardless of their actual grade level. In both the fall and the spring, students' executive function skills were assessed with the same measures fielded in kindergarten and first grade,[3] the *DCCS* and Numbers Reversed. Finally, children's height and weight were measured again in both fall and spring.

All children, regardless of home language, were administered the full assessment battery in English. By the spring of first grade, nearly all children had passed the screener, so it was not necessary to present the language screener or any of the assessments in a language other than English in the second-grade collections. In addition, the set of EBRS items was not administered, as the relatively low item difficulty of the set was not appropriate at the second-grade level. Exhibit 5-3 illustrates the administration of the second-grade assessments for all children, with the domains listed in order of administration.

Exhibit 5-3. Routing path for the direct child assessment in the ECLS-K:2011 second-grade year


Administer the assessments in English:
Reading
Mathematics
Science
Executive function[1]
Height and weight

[1] Two measures of executive function were administered. The *DCCS* was administered first, followed by Numbers Reversed.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2012 and spring 2013.

[3] Though the *Dimensional Change Card Sort (DCCS)* was fielded in each year of the study, the mode of administration changed from a physical card sort in kindergarten and first grade to a computerized version beginning with the second-grade data collections. The *DCCS* is discussed in more detail in chapter 6.

5.2 Scoring the National Assessment

This section presents information about the assessment scores developed for the kindergarten, first-grade, and second-grade 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 indicate 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 IRT Assumptions

In order to confirm that IRT was an appropriate estimation tool to use for scoring data from each of the assessments, confirmation of the multiple IRT assumptions discussed in chapter 3 (unidimensionality, monotonicity, and speededness) was performed. To confirm the assumption of unidimensionality, component analyses were run in each domain for each assessment year to determine if, indeed, the assessment for each domain was measuring a single, dominant component.

For reading, the kindergarten and first-grade component analyses showed a large single component but with second and third components that represented higher than expected percentages for unidimensionality. In second grade, the reading assessment showed a large single component with second and third components at the expected levels for unidimensionality. For math, the component analyses for each grade showed a large single component but with second and third components that represented higher than expected percentages for unidimensionality. For science, the kindergarten and first-grade component analyses showed a large single component followed by expected levels of second and third components for unidimensionality, while the second-grade component analysis showed a large single component but with second and third components higher than expected for unidimensionality. For the SERS assessment, the component analysis using data from the kindergarten year clearly showed a strong single component. Factor analyses were not performed for the SERS first-grade data as the sample sizes were too small for such analyses. Tables 5-1 through 5-3 show the percentage of the variance that each component in each domain explained for the kindergarten, first-grade, and second-grade data collections, respectively.

For those assessments with higher than expected second and third components, a review of component loadings was performed to determine whether the components were representing content components or simply differences in difficulty levels. In general, the greater the difficulty among the items in the test, the higher the rank of the matrix of inter-correlations: that is, differences in difficulty are represented in the factorial configuration as additional factors (Ferguson 1941). In reading, the component structure of the items with loadings was largely consistent with children's acquisition of early reading skills. Reading experts reviewed the component structure and, in particular, the content and difficulty of the items loading onto each component, and they determined that the components seem to correspond relatively well to the sophistication of reading acquisition from one phase to the next and the associated skills one might expect a child to possess. That is, 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 rather than content. As in reading, the component structure indicates multiple components based on increasing item difficulty, not content, and thus validates the assumption of unidimensionality. In science, items loaded on multiple components based on increasing item difficulty, even with the increasingly diverse content matter.

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. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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.

Table 5-2. Component analysis percentages by component by domain, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12

	Percentage of Component 1	Percentage of Component 2	Percentage of Component 3
Reading	10.42	5.36	3.44
Mathematics	10.23	6.53	4.32
Science	12.05	4.97	3.38

NOTE: Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011 and spring 2012.

Table 5-3. Component analysis percentages by component by domain, ECLS-K:2011 fall and spring second-grade data collections: School year 2012–13

	Percentage of Component 1	Percentage of Component 2	Percentage of Component 3
Reading	19.27	7.79	4.88
Mathematics	11.26	6.64	3.74
Science	9.98	5.43	3.14

NOTE: Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012 and spring 2013.

Unlike the tests for unidimensionality, there were not any empirical measures used to confirm monotonicity and speededness. With monotonicity, the probability of a correct response increases with increasing ability level. Monotonicity was confirmed by visual inspection of the item characteristic curves illustrating the IRT results. Since only items exhibiting monotonicity are selected for the national assessments, it was not anticipated that monotonicity would be an issue. Review of the item characteristic curves confirmed this assumption.

The assumption of speededness, 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, was confirmed by inspection of the model fit to data of the item characteristics curves and by examining the percentages of children not reaching the end of the assessment.

Poor model fit across rounds may be an indicator of speededness. Only items with the best model fit were selected for the national assessments, so speededness was not anticipated, and was not evident in review of the item characteristic curves.

Percentages of children not reaching the end of the assessment were reviewed for each data collection round (fall kindergarten, spring kindergarten, fall first grade, spring first grade, fall second grade and spring second grade), in each domain (reading, mathematics, and science), for each form (routing, low, medium, and high). For all forms that did not include discontinue rules, approximately 95–100 percent of the sample responded to the last item in each form. For those forms with discontinue rules, the number of student responses for the last item in the form was not useful in determining percentage complete. Thus, for each form, the item that was the last item administered, prior to any application of discontinue rules, was reviewed. Similar percentages complete (95–100 percent) were observed, for all but the middle reading form in fall first grade, where the percentage complete was 92 percent. Based on these results, speededness was not exhibited in any of the assessment forms.

5.2.2 Analysis of Differential Item Functioning (DIF)

Before scores were computed for each of the different subject area assessments in each year, 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 in the national data collections. (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.)

5.2.2.1 Kindergarten Rounds

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 5-4 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,[4] 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

[4] 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.

were not used for the SERS due to the limited number of children with SERS data in the fall and spring of kindergarten.

Table 5-4. 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 5-5 summarizes the results of the DIF analysis for 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 observable bias. DIF procedures were also used to analyze the spring kindergarten science assessment. None of the science items exhibited DIF in the spring of kindergarten.

Table 5-5. 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.

5.2.2.2 First-Grade Rounds

As in the kindergarten rounds, the M-H DIF and 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 5-6 summarizes the results of the M-H DIF and P-DIF analyses for all reading items for both first-grade rounds. Both C-level M-H DIF and P-DIF against one or more race/ethnicity focal groups were observed for five items. Two items were found to favor the focal group. Upon review, all items were retained for a variety of reasons, as stated above. DIF procedures were not used for the SERS due to the limited number of children with SERS data in the fall and spring of first grade.

Table 5-6. Reading assessment differential item functioning, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	0	5	1
Number of DIF items favoring focal group	0	0	2	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 2011 and spring 2012.

Table 5-7 summarizes the results of the DIF analysis for the fall and spring first-grade rounds combined in mathematics. Four items exhibited both C-level DIF and P-DIF against one or more race/ethnicity focal groups. Upon review of the items, no items were removed from scoring since none were determined to exhibit any observable bias.

Table 5-7. Mathematics assessment differential item functioning, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference groups	0	3	2	1
Number of DIF items favoring focal group	0	0	0	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 2011 and spring 2012.

DIF procedures were also used to analyze the first-grade science assessment. Table 5-8 summarizes the results for the fall and spring rounds combined in science. One item exhibited both C-level DIF and P-DIF against one race-ethnicity focal group. Upon its review, the item was retained since it was determined that the item did not exhibit any observable bias.

Table 5-8. Science assessment differential item functioning, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference groups	0	0	0	1
Number of DIF items favoring focal group	0	0	0	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). 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 2011 and spring 2012.

5.2.2.3 Second-Grade Rounds

As in the prior rounds, the M-H and 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. Table 5-9 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 race/ethnicity focal group were observed for one item. Upon review, this item was retained since bias was not indicated.

Table 5-9. Reading assessment differential item functioning, ECLS-K:2011 fall and spring second-grade data collections: School year 2012–13

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	0	0	1
Number of DIF items favoring focal group	0	0	0	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). 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 2012 and spring 2013.

Table 5-10 summarizes the results of the DIF analysis for the fall and spring second-grade rounds combined in mathematics. One item exhibited borderline C-level DIF and P-DIF against females, but it was retained for scoring since it was not determined to exhibit any observable bias.

Table 5-10. Mathematics assessment differential item functioning, ECLS-K:2011 fall and spring second-grade data collections: School year 2012–13

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference groups	1	0	0	0
Number of DIF items favoring focal group	0	0	0	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). 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 2012 and spring 2013.

DIF procedures were also used to analyze the second-grade science assessment. Table 5-11 summarizes the results of the fall and spring rounds combined in science. Two items exhibited both C-level DIF and P-DIF: one was a borderline DIF case against females, with the other against one race/ethnicity focal group. Upon review, these items were retained since it was determined they did not exhibit any observable bias.

Table 5-11. Science assessment differential item functioning, ECLS-K:2011 fall and spring second-grade data collections: School year 2012–13

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference groups	1	0	0	1
Number of DIF items favoring focal group	0	0	0	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). 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 2012 and spring 2013.

5.2.3 Assessment Score Reliability

Estimates of the reliability for each score (by domain by assessment year) are computed using the alpha coefficient for the number-right score and the reliability of the overall IRT ability estimate (Lord 1980) and presented in tables 5-17, 5-22, 5-27, and 5-34 below. Cronbach's alpha (Cronbach 1951), 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 the 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 Response Theory (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 is used to calculate scores for the ECLS-K:2011, as discussed in chapter 2, for two reasons. First, the study employed two-stage assessments[5] 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). Second, different assessment forms were administered in kindergarten, first grade, and second grade, also resulting in children being administered different sets of items across grades. However, by design there was a subset of items that was included in the assessments for more than one grade. These sets of common items (within grade and across grades) are 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 kindergarten science assessment and SERS) 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,[6] 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

[5] Two-stage assessments were administered in reading and mathematics in the kindergarten rounds and in reading, mathematics, and science in the first-grade and second-grade rounds.

[6] 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.

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, or different sets of items across grades.

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, science, or SERS) 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 are developed for each domain for each round in which an assessment in the domain is administered. The theta scores are reported on a metric ranging from -8 to 8, 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, on the condition that the scores are linked. 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 result to mean the child is doing better in reading than in mathematics). Gain scores may be calculated using any of the thetas available (from a single domain) for any round of data collection.

As described in chapter 3, both concurrent calibration and chain-linking were used to develop scores allowing for longitudinal measurement within and across grades. Within a given 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. The chain-linking approach is used to place the item parameters and ability (theta) estimates on the same scale *across* grades. Use of chain-linking allows for the calculation of theta for a given round that will not change based on later administrations of the assessments (which is not true for the scale scores, as described in the next section). Therefore, for any given child, the kindergarten, first-grade, and second-grade theta scores provided in subsequent data files will be the same as theta scores released in earlier data files, with one exception: the reading thetas provided in the base-year data file. As stated in section 3.5.2, after the kindergarten year data collection, the methodology used to calibrate and compute reading scores changed; therefore, the reading thetas reported in the base-year file are not the same as the kindergarten reading thetas provided in the files with later-round data. Any analysis involving kindergarten reading theta scores and reading theta scores from later rounds (e.g., an analysis looking at growth in reading knowledge and skills between the spring of kindergarten and the spring of first grade) should use the kindergarten reading theta scores from a data file released after the base year. The reading theta scores released in the kindergarten-year data file *are* appropriate for analyses involving only the kindergarten-round data; analyses conducted with data released in the base-year file are *not* incorrect since those analyses do not compare kindergarten scores to scores in later rounds that were computed differently. However, now that the recomputed kindergarten theta scores are available in the kindergarten-first grade and kindergarten-second grade data files, it is recommended that researchers conduct any new analyses with the recomputed kindergarten reading theta scores.

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, which assumes the precision of the scores is usually 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 (also known as "domain 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 unique questions for that domain in all rounds (Bock, Thissen, and Zimowski 1997).

To calculate the IRT-based overall scale score for each domain, for each assessment item a child's theta is used to predict a probability 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, mathematics, science, or SERS) 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 the fall and spring administrations in a given grade, 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 score at an earlier administration from the IRT scale score 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. Gain scores may be calculated using any of the scale scores available (from a single domain) for any round of data collection. However, the scale scores are only comparable across rounds within a single data file. In other words, the scale scores for a given domain in the kindergarten through second-grade data file are all comparable to one other, but they are not comparable to the scale scores for that domain reported in the base-year file or in subsequent files. Although the thetas remain the same for a given domain across rounds, the scale scores are recomputed for each file because the scale scores represent the estimated number correct for *all* items across *all* assessments administered; the total number of items in the pool expands each year as more difficult items are added to the assessments.

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 and first-grade rounds of data collection.[7] Raw number-right scores for the kindergarten and first-grade administrations of 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. The scores 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 in kindergarten, so all children have raw number-right scores for these two subtests in both kindergarten rounds of data collection. In first grade, the *pre*LAS items were administered only to children who spoke a language other than English at home who had not passed the screener in the most recent round in which they were assessed. Thus, scores for these sets of items are not available for all children in first grade.

A raw number-right score also is provided for children's performance on the set of 20 EBRS items in kindergarten and first grade. In kindergarten, the EBRS items were administered to all children as part of the reading assessment routing test, so all children have an EBRS score. As noted above in section 5.1, in first grade the EBRS items were administered only to children who did not pass the language screener and to children who were routed to the low- and middle-difficulty second-stage tests. Therefore, first-grade EBRS scores are only available for a subset of children. Additionally, for those children who were routed to the SERS in the kindergarten and first grade rounds, number-right scores are provided for the 10 items common to the EBRS and SERS. EBRS scores are not provided for the second-grade rounds of data collection because the reading assessment no longer included all the English basic reading skills items used to calculate the EBRS scores in the earlier rounds. SERS scores are not computed for the second-grade rounds because the Spanish reading assessment was not fielded in second grade.

5.3 Reading Assessment

5.3.1 Samples and Associated Statistics for the Kindergarten Rounds

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-

[7] Raw number-right scores were not computed beyond first grade since the assessments for which raw-number right scores are computed were not administered in second grade.

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 as easier items 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 5-12. There was no evidence of floor or ceiling effects (based on the low numbers of children with chance[8] 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.[9] Review of the classical item analysis *r*-biserials (see section 3.1 for information about classical item analysis) in both the fall and spring showed two items with slightly lower *r*-biserials than the ideal minimum of .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.

[8] The chance score is defined as the lowest possible score on a test. For tests that include only constructed response items, the chance score computed using IRT is approximately zero. But for tests including multiple-choice items, there is the possibility of guessing. Therefore, for tests that include multiple-choice items, the chance score computed using IRT is approximately the sum of the IRT guessing parameters for the multiple-choice items on the test. On most tests, if two scores are both below the chance score, the difference between those scores provides very little information about the difference between the two test takers who earned those scores.

[9] Individual items were counted toward the scoreability threshold. Polytomous item sets were not counted as single items; each individual item in the set counted toward the threshold. For example, although a reading passage may have three associated assessment items that are scored together as one polytomous item, those three items are counted separately when determining whether a child has scoreable data.

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

	Fall kindergarten		Spring kindergarten	
Characteristic	Number	Percent	Number	Percent
Total sample size	15,790	100	17,210	100
Number of children with responses to fewer than 10 items	120	1	30	#
Number of children with scoreable data	15,670	99	17,190	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	#	10	#
Number and percent of children routed to low form	3,040	19	420	2
Number and percent of children routed to middle form	10,990	70	9,980	58
Number and percent of children routed to high form	1,280	8	6,610	38
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	30	#	50	#

# Rounds to zero.

NOTE: The unweighted *n* is the rounded number of cases. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted. Percentages may not sum to 100 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; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.3.2 Samples and Associated Statistics for the First-Grade Rounds

The first-grade reading assessment consisted of 30 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 14, 21, and 15 items, respectively. Prior to the administration of the low- and middle-difficulty second-stage forms, children first received 18 (of 20) items that contributed to the calculation of the EBRS score.[10,11] As with the kindergarten assessment, discontinue rules were employed within the second-stage tests 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 as easier items 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 for the first-grade rounds are shown in table 5-13. There was no evidence of floor or ceiling effects on the reading assessment during the fall or spring first-grade rounds. Review of the classical item analysis *r*-biserials in both the fall and spring showed three very easy items with lower *r*-biserials than the ideal minimum of 0.3.

[10] The other two EBRS items were administered as part of the *pre*LAS to the children administered those subtests, or as two additional items for the children not administered the *pre*LAS subtests.

[11] In the fall and spring first-grade administrations of the reading assessment, the EBRS items were administered only to children whose performance on the routing items of the reading assessment routed them to the low- or middle-difficulty second-stage test. Children who were routed to the highest-difficulty second-stage test did not receive the EBRS items. Therefore, only a subsample of children have EBRS raw-number right scores in first grade.

Table 5-13. First-grade reading assessment samples, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12

	Fall first grade		Spring first grade	
Characteristic	Number	Percent	Number	Percent
Total sample size	5,230	100	15,130	100
Number of children with responses to fewer than 10 items	40	1	20	#
Number of children with scoreable data	5,190	100	15,120	100
Number and percent of children with responses only for the router	#	#	10	#
Number and percent of children with responses only for the EBRS items	0	0	20	#
Number and percent of children routed to low form	1,990	38	1,550	10
Number and percent of children routed to middle form	2,610	50	7,850	52
Number and percent of children routed to high form	600	11	5,690	38
Number and percent of children with a perfect score: router + high form	0	0	#	#
Number and percent of children with a chance score or below: router + EBRS + low form	#	#	10	#

# Rounds to zero.

NOTE: The unweighted *n* is the rounded number of cases. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted. Percentages may not sum to 100 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 2011 and spring 2012; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.3.3 Samples and Associated Statistics for the Second-Grade Rounds

The second-grade reading assessment consisted of 29 routing items followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 21, 22, and 22 items, respectively. Discontinue rules were employed as in prior rounds.

The total number of children who were administered the reading assessment and the assessment's associated statistics are shown in table 5-14. In the fall administration, about 3 percent of the children scored at or below chance, resulting in a borderline case of a floor effect for the subsample.

There was no evidence of a floor effect in the spring administration, and no evidence of ceiling effects in either the fall or spring. Review of the classical item analysis *r*-biserials in both the fall and spring showed three items with slightly lower *r*-biserials than the ideal minimum of .3: one item that was very difficult for the majority of the sample and two items that were very easy for the majority of the sample.

Table 5-14. Second-grade reading assessment samples, ECLS-K:2011 fall and spring second-grade data collections: School year 2012–13

	Fall second grade		Spring second grade	
Characteristic	Number	Percent	Number	Percent
Total sample size	4,740	100	13,850	100
Number of children with responses to fewer than 10 items	10	#	10	#
Number of children with scoreable data	4,730	100	13,840	100
Number and percent of children with responses only for the router	0	0	0	0
Number and percent of children routed to low form	2,410	51	4,000	29
Number and percent of children routed to middle form	1,810	38	6,380	46
Number and percent of children routed to high form	510	11	3,460	25
Number and percent of children with a perfect score: router + high form	0	0	0	0
Number and percent of children with a chance score or below: router + low form	150	3	160	1

# Rounds to zero.

NOTE: The unweighted *n* is the rounded number of cases. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted. Percentages may not sum to 100 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 2012 and spring 2013; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten--Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.3.4 Score Statistics

Table 5-15 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 120 unique items administered in the kindergarten, first-grade, and second-grade rounds. Table 5-16 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. These raw number-right scores are integers based on the total number of items administered in each subset and are reported only for the kindergarten and first-grade rounds because the *pre*LAS and EBRS item sets were not administered in second grade. Both the IRT-based scores and the raw number-right scores are calculated for all children with scoreable reading assessment data.[12]

Table 5-15. Reading assessment statistics, by IRT-based score, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, and fall and spring second-grade data collections: School years 2010–11, 2011–12, and 2012–13

Variable	Description	*n*	Range of possible values	Weighted mean	Weighted standard deviation
X1RTHETK2	X1 READING THETA	15,670	-8.0–+8.0	-0.56	0.844
X2RTHETK2	X2 READING THETA	17,190	-8.0–+8.0	0.44	0.774
X3RTHETK2	X3 READING THETA	5,190	-8.0–+8.0	0.87	0.778
X4RTHETK2	X4 READING THETA	15,120	-8.0–+8.0	1.59	0.753
X5RTHETK2	X5 READING THETA	4,730	-8.0–+8.0	1.83	0.654
X6RTHETK2	X6 READING THETA	13,840	-8.0–+8.0	2.19	0.630
X1RSETHK2	X1 READING STD ERR OF THETA	15,670	0.0–+6.0	0.36	0.092
X2RSETHK2	X2 READING STD ERR OF THETA	17,190	0.0–+6.0	0.26	0.077
X3RSETHK2	X3 READING STD ERR OF THETA	5,190	0.0–+6.0	0.23	0.057
X4RSETHK2	X4 READING STD ERR OF THETA	15,120	0.0–+6.0	0.22	0.050
X5RSETHK2	X5 READING STD ERR OF THETA	4,730	0.0–+6.0	0.20	0.035
X6RSETHK2	X6 READING STD ERR OF THETA	13,840	0.0–+6.0	0.22	0.042
X1RSCALK2	X1 READING IRT SCALE SCORE	15,670	0.0–+120.0	46.70	11.284
X2RSCALK2	X2 READING IRT SCALE SCORE	17,190	0.0–+120.0	61.02	13.276
X3RSCALK2	X3 READING IRT SCALE SCORE	5,190	0.0–+120.0	68.94	15.474
X4RSCALK2	X4 READING IRT SCALE SCORE	15,120	0.0–+120.0	84.21	15.520
X5RSCALK2	X5 READING IRT SCALE SCORE	4,730	0.0–+120.0	89.14	13.609
X6RSCALK2	X6 READING IRT SCALE SCORE	13,840	0.0–+120.0	96.25	12.100

NOTE: The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (1 for the fall kindergarten round, 2 for the spring kindergarten round, 3 for the fall first-grade round, 4 for the spring first-grade round, 5 for the fall second-grade round, and 6 for the spring second-grade round). X1 and X2 estimates weighted by W1C0; X3 estimates weighted by W3CF3P_30; X4 estimates weighted by W4CS4P_20; X5 estimates weighted by W6CF6P_2A0; and X6 estimates weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. 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) Restricted-Use Kindergarten--Second Grade Data File and Electronic Codebook (NCES 2015-050).

[12] Plots of the IRT theta and scale score distributions are provided in appendix D.

Table 5-16. Reading assessment statistics, by raw number-right score, ECLS-K:2011 fall and spring kindergarten and fall and spring first-grade data collections: School years 2010–11 and 2011–12

Variable	Description	*n*	Range of possible values	Weighted mean	Weighted standard deviation
X1PLSS	X1 PRELAS SIMON SAYS SCORE	15,780	0–10	9.18	1.754
X2PLSS	X2 PRELAS SIMON SAYS SCORE	17,220	0–10	9.60	1.120
X3PLSS	X3 PRELAS SIMON SAYS SCORE	110	0–10	6.44	2.917
X4PLSS	X4 PRELAS SIMON SAYS SCORE	100	0–10	7.27	3.068
X1PLART	X1 PRELAS ART SHOW SCORE	15,780	0–10	9.26	1.705
X2PLART	X2 PRELAS ART SHOW SCORE	17,220	0–10	9.54	1.274
X3PLART	X3 PRELAS ART SHOW SCORE	110	0–10	4.36	2.383
X4PLART	X4 PRELAS ART SHOW SCORE	100	0–10	6.15	2.870
X1PLTOT	X1 PRELAS TOTAL SCORE	15,780	0–20	18.43	3.184
X2PLTOT	X2 PRELAS TOTAL SCORE	17,220	0–20	19.14	2.178
X3PLTOT	X3 PRELAS TOTAL SCORE	110	0–20	10.80	4.879
X4PLTOT	X4 PRELAS TOTAL SCORE	100	0–20	13.42	5.170
X1EBRSTOT	X1 EBRS TOTAL NUMBER RIGHT	15,740	0–20	13.18	4.424
X2EBRSTOT	X2 EBRS TOTAL NUMBER RIGHT	17,200	0–20	17.06	2.976
X3EBRSTOT	X3 EBRS TOTAL NUMBER RIGHT	4,630	0–20	14.88	2.313
X4EBRSTOT	X4 EBRS TOTAL NUMBER RIGHT	9,430	0–20	15.37	2.015

NOTE: The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (1 for the fall kindergarten round, 2 for the spring kindergarten round, 3 for the fall first-grade round, and 4 for the spring first-grade round). X1 and X2 estimates weighted by W1C0; X3 estimates weighted by W3CF3P_30; X4 estimates weighted by W4CS4P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. EBRS = English 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) Restricted-Use Kindergarten--Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.3.5 Reliabilities

Table 5-17 presents the reliability statistics for the reading assessment scores. The reliabilities shown in table 5-17 are typical and adequate for tests with these numbers of items.

Table 5-17. Reading assessment reliabilities, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, and fall and spring second-grade data collections: School years 2010–11, 2011–12, and 2012–13

Score	Number of items	*n*	Weighted reliability
Fall kindergarten			
IRT-based scores	54[1]	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	54[1]	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	17,200	.97
Fall first grade			
IRT-based scores	62	5,190	.95
*pre*LAS Simon Says raw number-right score	10	110	.83
*pre*LAS Art Show raw number-right score	10	110	.89
*pre*LAS total raw number-right score	20	110	.91
EBRS raw number-right score	20	4,630	.94
Spring first grade			
IRT-based scores	62	15,120	.93
*pre*LAS Simon Says raw number-right score	10	100	.88
*pre*LAS Art Show raw number-right score	10	100	.89
*pre*LAS total raw number-right score	20	100	.92
EBRS raw number-right score	20	9,430	.99
Fall second grade			
IRT-based scores	76	4,730	.93
Spring second grade			
IRT-based scores	76	13,840	.91

[1] The switch to a polytomous scoring methodology in reading (described in section 3.5.2) reduced the total number of items from 83 to 54, since multiple items were collapsed into polytomous sets.

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 defined as the standard deviation of the posterior estimate) to total variance (across the sample). The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. IRT = item response theory. EBRS = English 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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten—Second-Grade Data File and Electronic Codebook (NCES 2015-050).

5.4 Spanish Early Reading Skills (SERS)

5.4.1 Samples and Associated Statistics for the Kindergarten Rounds

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 5-18. 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 the 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 5-18. 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	0	0
Number and percent of children with a chance score	10	2	0	0

NOTE: Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. The unweighted *n* is the rounded number of cases with a valid score. Percentages are unweighted and calculated using unrounded sample sizes. 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. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten--Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.4.2 Samples and Associated Statistics for the First-Grade Rounds

The samples and associated statistics for the fall and spring first-grade administrations of the SERS are shown in table 5-19. Analysis of any floor or ceiling effects in the first-grade SERS assessment is not possible due to the small number of children who were administered the SERS in each first-grade round. Five of the items in the fall administration had lower *r*-biserials than ideal, likely a result of the small sample.

Table 5-19. First-grade SERS assessment samples, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12

	Fall first grade		Spring first grade	
Characteristics	Number	Percent	Number	Percent
Total sample size	30	100	20	100
Number and percent of children with a perfect score	#	6	0	0
Number and percent of children with a chance score	0	0	0	0

# Rounds to zero.

NOTE: The unweighted *n* is the rounded number of cases. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. The unweighted *n* is the rounded number of cases with a valid score. Percentages are unweighted and calculated using unrounded sample sizes. 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 2011 and spring 2012; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten--Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.4.3 Score Statistics

Table 5-20 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 and first-grade rounds. Table 5-21 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. Note that for both the IRT-based and raw number-right SERS scores, the mean spring first-grade scores are lower, and the standard deviations greater, than those at fall first-grade. Examination of the data showed one student with a low estimated ability relative to the rest of the sample. The mean score of the sample with this child’s score removed is higher than that of the fall first-grade sample, with a standard deviation more aligned with the other rounds.

Table 5-20. SERS assessment statistics by IRT-based score, ECLS-K:2011 fall and spring kindergarten and fall and spring first-grade data collections: School years 2010–11 and 2011–12

Variable	Description	*N*	Range of possible values	Weighted mean	Weighted standard deviation
X1SERSTHK1	X1 SERS THETA	310	-8.0–+8.0	-0.41	0.852
X2SERSTHK1	X2 SERS THETA	150	-8.0–+8.0	0.68	0.667
X3SERSTHK1	X3 SERS THETA	30	-8.0–+8.0	0.90	0.744
X4SERSTHK1	X4 SERS THETA	20	-8.0–+8.0	0.27	2.004
X1SERSSEK1	X1 SERS STD ERR OF THETA	310	0.0–6.0	0.39	0.142
X2SERSSEK1	X2 SERS STD ERR OF THETA	150	0.0–6.0	0.27	0.077
X3SERSSEK1	X3 SERS STD ERR OF THETA	30	0.0–6.0	0.32	0.128
X4SERSSEK1	X4 SERS STD ERR OF THETA	20	0.0–6.0	0.49	0.341
X1SERSSCK1	X1 SERS IRT SCALE SCORE	310	0.0–31.0	12.75	5.343
X2SERSSCK1	X2 SERS IRT SCALE SCORE	150	0.0–31.0	20.83	5.592
X3SERSSCK1	X3 SERS IRT SCALE SCORE	30	0.0–31.0	23.43	6.011
X4SERSSCK1	X4 SERS IRT SCALE SCORE	20	0.0–31.0	21.67	10.455

NOTE: The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (1 for the fall kindergarten round, 2 for the spring kindergarten round, 3 for the fall first-grade round, and 4 for the spring first-grade round). X1 and X2 estimates weighted by W1C0; X3 estimates weighted by W3CF3P_30; X4 estimates weighted by W4CS4P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. SERS = Spanish early reading skills. 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) Restricted-Use Kindergarten--Second Grade Data File and Electronic Codebook (NCES 2015-050).

Table 5-21. SERS and EBRS common item assessment statistics by raw number-right score, ECLS-K:2011 fall and spring kindergarten and fall and spring first-grade data collections: School years 2010–11 and 2011–12

Variable	Description	*n*	Range of possible values	Weighted mean	Weighted standard deviation
X1EBRSCM	X1 EBRS NUMBER RIGHT COMMON ITEMS	340	0–10	3.21	2.583
X2EBRSCM	X2 EBRS NUMBER RIGHT COMMON ITEMS	150	0–10	4.13	3.129
X3EBRSCM	X3 EBRS NUMBER RIGHT COMMON ITEMS	60	0–10	5.63	2.628
X4EBRSCM	X4 EBRS NUMBER RIGHT COMMON ITEMS	40	0–10	6.16	2.159
X1SERSCM	X1 SERS NUMBER RIGHT COMMON ITEMS	320	0–10	4.72	2.992
X2SERSCM	X2 SERS NUMBER RIGHT COMMON ITEMS	150	0–10	8.05	2.037
X3SERSCM	X3 SERS NUMBER RIGHT COMMON ITEMS	30	0–10	8.46	2.157
X4SERSCM	X4 SERS NUMBER RIGHT COMMON ITEMS	20	0–10	7.13	3.761

NOTE: The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (1 for the fall kindergarten round, 2 for the spring kindergarten round, 3 for the fall first-grade round, and 4 for the spring first-grade round). X1 and X2 estimates weighted by W1C0; X3 estimates weighted by W3CF3P_30; X4 estimates weighted by W4CS4P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. EBRS = English basic reading skills. SERS = Spanish early reading skills. *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) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

5.4.4 Reliabilities

Table 5-22 presents the reliability statistics for the scores of the SERS assessment for the kindergarten and first-grade rounds. All scores, except the EBRS common raw number-right score, have reliabilities that are typical and adequate for assessments with this many items. The EBRS common raw number-right score has a lower reliability than generally acceptable due to the low number of items contributing to the score combined with the lack of variability in the Spanish-speaking children's abilities in English.

Table 5-22. SERS assessment reliabilities, ECLS-K:2011 fall and spring kindergarten and fall and spring first-grade data collections: School years 2010–11 and 2011–12

Score	Number of items	*n*	Weighted 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
Fall first grade			
IRT-based scores	31	30	.91
EBRS common raw number-right score	10	60	.92
SERS common raw number-right score	10	30	.86
Spring first grade			
IRT-based scores	31	20	.99
EBRS common raw number-right score	10	40	.96
SERS common raw number-right score	10	20	.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 defined as the standard deviation of the posterior estimate) to total variance (across the sample). The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. IRT = item response theory. EBRS = English 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, spring 2011, fall 2011, and spring 2012; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

5.5 Mathematics Assessment

5.5.1 Samples and Associated Statistics for the Kindergarten Rounds

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 5-23 shows 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 (based on the low numbers of children with perfect or chance scores, respectively) were observed in the fall or spring. 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. However, due to the low number of observations, the *r*-biserial calculations for this small subset may be unreliable, suggesting that these results should not be taken as indicating that the items did not function well.

5.5.2 Samples and Associated Statistics for the First-Grade Rounds

The first-grade mathematics assessment consisted of 17 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 23, 28, and 29 items, respectively. Table 5-24 shows the total number of children administered the mathematics assessment and the assessment's associated statistics for the fall and spring first-grade rounds. No significant ceiling or floor effects were observed in the fall or spring. Classical item analysis results for the English mathematics administration showed five very easy items with *r*-biserials lower than ideal. These five items, although relatively easy for the majority of the sample, were needed to differentiate child ability at the low end of the ability spectrum. In a longitudinal study, ability estimates at the tails of the distribution are desired, thus, items that may be too easy (or too difficult) for the majority of the sample are still included to address estimation in the tails. The Spanish-translated mathematics assessment showed several items 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 5-23. Kindergarten mathematics 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	16,100	100	17,360	100
Number of children with responses to fewer than 10 items	510	3	220	1
Number of children with scoreable data	15,600	97	17,140	99
Number and percent of children with responses only for the router	0	0	#	#
Number and percent of children routed to low form	4,090	26	930	5
Number and percent of children routed to middle form	9,900	63	9,360	55
Number and percent of children routed to high form	1,600	10	6,850	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: The unweighted *n* is the rounded number of cases. Estimates are based on all children administered the mathematics assessment regardless of language of administration. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted.

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. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

Table 5-24. First-grade mathematics assessment samples, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12

	Fall first grade		Spring first grade	
Characteristics	Number	Percent	Number	Percent
Total sample size	5,230	100	15,130	100
Number of children with responses to fewer than 10 items	10	#	30	#
Number of children with scoreable data	5,220	100	15,100	100
Number and percent of children with responses only for the router	30	1	20	#
Number and percent of children routed to low form	760	15	650	4
Number and percent of children routed to middle form	3,340	64	6,900	46
Number and percent of children routed to high form	1,080	21	7,540	50
Number and percent of children with a perfect score: router + high form	0	0	#	#
Number and percent of children with chance score or below: router + low form	#	#	10	#

# Rounds to zero.
NOTE: The unweighted *n* is the rounded number of cases. Estimates are based on all children administered the mathematics assessment regardless of language of administration. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2011 and spring 2012; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.5.3 Samples and Associated Statistics for the Second-Grade Rounds

The second-grade mathematics assessment consisted of 20 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 28, 29, and 20 items, respectively. Table 5-25 shows the total number of children administered the mathematics assessment and the assessment's associated statistics for the fall and spring second-grade rounds. No significant ceiling or floor effects were observed in the fall or spring. Classical item analysis results for the English mathematics administration showed no items with *r*-biserials lower than ideal. As noted above in section 5.1, there was no Spanish administration of the mathematics assessment in second grade.

Table 5-25. Second-grade mathematics assessment samples, ECLS-K:2011 fall and spring second-grade data collections: School year 2012–13

	Fall second grade		Spring second grade	
Characteristics	Number	Percent	Number	Percent
Total sample size	4,740	100	13,850	100
Number of children with responses to fewer than 10 items	10	#	20	#
Number of children with scoreable data	4730	100	13,830	100
Number and percent of children with responses only for the router	0	0	#	#
Number and percent of children routed to low form	2,130	45	3,090	22
Number and percent of children routed to middle form	2,440	51	8,530	62
Number and percent of children routed to high form	170	4	2,220	16
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	10	#	9	#

# Rounds to zero.

NOTE: The unweighted *n* is the rounded number of cases. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2012 and spring 2013. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

5.5.4 Score Statistics

Only IRT-based scores were produced for the mathematics assessment. Table 5-26 presents summary statistics for the mathematics thetas, the standard errors of measurement (*SEM*) of thetas, and the IRT scale scores, using the 113 unique items administered in the kindergarten, first-grade, and second-grade rounds.[13]

[13] Plots of the IRT theta and scale score distributions are provided in appendix D.

Table 5-26. Mathematics assessment statistics by IRT-based score, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, and fall and spring second-grade data collections: School years 2010–11, 2011–12, and 2012–13

Variable	Description	*n*	Range of possible values	Weighted mean	Weighted standard deviation
X1MTHETK2	X1 MATH THETA	15,600	-8.0–+8.0	-0.52	0.932
X2MTHETK2	X2 MATH THETA	17,140	-8.0–+8.0	0.42	0.776
X3MTHETK2	X3 MATH THETA	5,220	-8.0–+8.0	0.91	0.821
X4MTHETK2	X4 MATH THETA	15,100	-8.0–+8.0	1.65	0.840
X5MTHETK2	X5 MATH THETA	4,730	-8.0–+8.0	1.91	0.817
X6MTHETK2	X6 MATH THETA	13,830	-8.0–+8.0	2.45	0.802
X1MSETHK2	X1 MATH STD ERR OF THETA	15,600	0.0–6.0	0.36	0.101
X2MSETHK2	X2 MATH STD ERR OF THETA	17,140	0.0–6.0	0.29	0.064
X3MSETHK2	X3 MATH STD ERR OF THETA	5,220	0.0–6.0	0.28	0.048
X4MSETHK2	X4 MATH STD ERR OF THETA	15,100	0.0–6.0	0.28	0.037
X5MSETHK2	X5 MATH STD ERR OF THETA	4,730	0.0–6.0	0.29	0.043
X6MSETHK2	X6 MATH STD ERR OF THETA	13,830	0.0–6.0	0.28	0.035
X1MSCALK2	X1 MATH IRT SCALE SCORE	15,600	0.0–113.0	31.32	11.243
X2MSCALK2	X2 MATH IRT SCALE SCORE	17,140	0.0–113.0	44.86	12.217
X3MSCALK2	X3 MATH IRT SCALE SCORE	5,220	0.0–113.0	53.35	14.660
X4MSCALK2	X4 MATH IRT SCALE SCORE	15,100	0.0–113.0	66.82	15.187
X5MSCALK2	X5 MATH IRT SCALE SCORE	4,730	0.0–113.0	71.70	14.540
X6MSCALK2	X6 MATH IRT SCALE SCORE	13,830	0.0–113.0	81.04	13.587

NOTE: The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (1 for the fall kindergarten round, 2 for the spring kindergarten round, 3 for the fall first-grade round, 4 for the spring first-grade round, 5 for the fall second-grade round, and 6 for the spring second-grade round). X1 and X2 estimates weighted by W1C0; X3 estimates weighted by W3CF3P_30; X4 estimates weighted by W4CS4P_20; X5 estimates weighted by W6CF6P_2A0; and X6 estimates weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. IRT = item response theory. *SEM* = standard error of measurement. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

5.5.5 Reliabilities

Table 5-27 presents reliability statistics for the scores of the fall and spring kindergarten, fall and spring first-grade, and fall and spring second-grade mathematics assessments (calculated in the same way as the reading reliability statistics, described in section 5.3.5). The reliabilities shown in table 5-14 are typical and adequate for tests with this number of items.

Table 5-27. Mathematics assessment reliabilities, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, and fall and spring second-grade data collections: School years 2010–11, 2011–12, and 2012–13

Score	Number of items	*n*	Weighted reliability
Fall kindergarten			
IRT-based scores	75	15,600	.92
Spring kindergarten			
IRT-based scores	75	17,140	.94
Fall first grade			
IRT-based scores	80	5,220	.93
Spring first grade			
IRT-based scores	80	15,100	.93
Fall second grade			
IRT-based scores	113	4,730	.92
Spring second grade			
IRT-based scores	113	13,830	.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 defined as the standard deviation of the posterior estimate) to total variance (across the sample). The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. 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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

5.5.6 Comparability of the English and Spanish Mathematics Tests

The mathematics assessment was translated from English into Spanish for administration in the kindergarten and first-grade rounds of data collection. In these grades, 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. When an assessment item shows DIF by the language of administration, it could be an indication that children with equal mathematics ability demonstrated differential performance attributable to the language in which the item was administered. Lack of DIF can be taken as an indication 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 for the kindergarten rounds showed that none of the items exhibited significant DIF, as defined in section 3.4.[14] The sample sizes were too small in the first-grade rounds to perform DIF procedures. However, given the lack of DIF observed between the items when administered in English or Spanish in kindergarten, and since many of the same items were administered in the first-grade rounds, data from the English and Spanish administrations were calibrated and scored as the same assessment by grade, regardless of the language of administration, rather than scoring them as separate assessments.

Tables 5-28 and 5-29 show the samples and associated statistics of the mathematics assessment by language of administration for the kindergarten and first-grade rounds, respectively.

Table 5-28. 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	17,360	100	150	100
Number of children with responses to fewer than 10 items	510	3	#	1	220	1	#	1
Number of children with scoreable data	15,600	97	310	99	17,140	99	150	99

NOTE: The unweighted *n* is the rounded number of cases. Percentages are unweighted.
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. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

[14] 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 when the fall and spring data were combined. With smaller sample sizes, artificial statistical DIF may result, which is why the minimum sample size is required. Therefore, with the lower number of observations, and still the non-existence of DIF, the analyses are considered valid.

Table 5-29. First-grade mathematics assessment samples, by language of assessment, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12

	Fall kindergarten				Spring kindergarten			
	English		Spanish		English		Spanish	
Characteristic	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Total sample size	5,200	100	30	100	15,120	100	20	100
Number of children with responses to fewer than 10 items	10	#	0	0	30	#	0	0
Number of children with scoreable data	5,190	100	30	100	15,090	100	20	100

# Rounds to zero.

NOTE: The unweighted *n* is the rounded number of cases. Percentages are unweighted.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2011 and spring 2012. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

5.6 Science Assessment

5.6.1 Samples and Associated Statistics in Spring Kindergarten

The spring kindergarten science assessment included a total of 20 items administered to all children who were routed to the assessment. Table 5-30 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 (based on the low numbers of children with perfect or chance scores, respectively) were observed for the spring kindergarten science test. The classical item analysis results showed one difficult item with an *r*-biserial slightly lower than ideal.

Table 5-30. Kindergarten science assessment sample: School year 2010–11

Characteristics	Number	Percent
Total sample size	17,220	100
Number of children with responses to fewer than 10 items	280	2
Number of children with scoreable data	16,940	98
Number and percent of children with a perfect score	20	#
Number and percent of children with chance score or below	260	2

# Rounds to zero.
NOTE: The unweighted *n* is the rounded number of cases. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted.
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. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

5.6.2 Samples and Associated Statistics in the First-Grade Rounds

The first-grade science assessment consisted of 15 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 14, 15, and 15 items, respectively. Table 5-31 shows the total number of children administered the science assessment and the assessment's associated statistics for the fall and spring first-grade rounds. No significant ceiling or floor effects were observed in the fall or spring. Classical item analysis results for the science administrations showed one item for which the correct response was selected by children with only a slightly higher average ability than those selecting the other options and four very easy items with *r*-biserials lower than ideal. All items were retained. The four relatively easy items for the majority of the sample were needed to differentiate child ability at the low end of the ability spectrum. In a longitudinal study, ability estimates at the tails of the distribution are desired, thus, items that may be too easy (or too difficult) for the majority of the sample are still included to address estimation in the tails.

During the analysis of the first-grade science data, it was determined that an error had been made in the routing specifications programmed for the first-grade science assessment program. During the administration of the assessment for each child, the computer application is programmed to calculate a total routing score to determine which of the second-stage tests the child should be administered. One of the router items was not scored properly in determining the total routing score, and thus routing to the second-stage form was affected in the following ways:

- Cases for which this router item was erroneously scored incorrect (that is, the child actually provided a correct answer) and that had a calculated routing score one point below the cut-score were routed to a lower second-stage form then they should have been. Had the item been appropriately scored as correct, the router score would have been one point higher (that is, at the cut-score), and the cases would have been routed to a second-stage form one level of difficulty higher than the second-stage form they were actually administered. Specifically, a child with a calculated routing score of 4 who would have had a score of 5 if this item had appropriately been scored as correct was routed to the low form but should have been routed to the middle form. Similarly, a child with a calculated routing score of 9 who would have had a score of 10 if this item had appropriately been scored as incorrect was routed to the middle form but should have been routed to the high form.

- Cases for which this router item was erroneously scored correct (that is, the child actually provided an incorrect answer) and that had a calculated routing score at the cut-score were routed to a higher second-stage form then they should have been. Had the item been appropriately scored as incorrect, the router score would have been one point lower (that is, one point below the cut score), and the cases would have been routed to a second-stage form one level of difficulty lower than the second-stage form they were actually administered. Specifically, a child with a calculated routing score of 5 who would have had a score of 4 if this item had appropriately been scored as incorrect was routed to the middle form but should have been routed to the low form. Similarly, a child with a calculated routing score of 10 who would have had a score of 9 if this item had appropriately been scored as incorrect was routed to the high form but should have been routed to the middle form.

Approximately 2,900 children in the first-grade sample were affected by this error, although review of the *SEM* of theta did not show any systematic differences for children who were and were not routed correctly. This finding is expected; with the overlap in items across adjacent second-stage forms, children misrouted due to this error were still presented many items appropriate for their ability levels.

Table 5-31. First-grade science assessment samples, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12

	Fall first grade		Spring first grade	
Characteristics	Number	Percent	Number	Percent
Total sample size	5,230	100	15,130	100
Number of children with responses to fewer than 10 items	50	1	60	#
Number of children with scoreable data	5,180	100	15,070	100
Number and percent of children with responses only for the router	0	0	#	#
Number and percent of children routed to low form	840	16	1,060	7
Number and percent of children routed to middle form	3,340	64	8,610	57
Number and percent of children routed to high form	1,010	19	5,400	36
Number and percent of children with a perfect score: router + high form	0	0	0	0
Number and percent of children with chance score or below: router + low form	20	#	30	#

# Rounds to zero.

NOTE: The unweighted *n* is the rounded number of cases. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2011 and spring 2012; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second-Grade Data File and Electronic Codebook (NCES 2015-050).

5.6.3 Samples and Associated Statistics in the Second-Grade Rounds

The second-grade science assessment consisted of 19 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 15, 15, and 16 items, respectively. Table 5-32 shows the total number of children administered the science assessment and the assessment's associated statistics for the fall and spring second-grade rounds. No significant ceiling or floor effects were observed in the fall or spring. Classical item analysis results for the science administrations showed one very easy item with an *r*-biserial lower than ideal.

Table 5-32. Second-grade science assessment samples, ECLS-K:2011 fall and spring second-grade data collections: School year 2012–13

Characteristics	Fall second grade		Spring second grade	
	Number	Percent	Number	Percent
Total sample size	4,740	100	13,850	100
Number of children with responses to fewer than 10 items	10	#	30	#
Number of children with scoreable data	4,720	100	13,820	100
Number and percent of children with responses only for the router	#	#	0	0
Number and percent of children routed to low form	2,300	49	4,120	30
Number and percent of children routed to middle form	2,010	43	6,960	50
Number and percent of children routed to high form	420	9	2,740	20
Number and percent of children with a perfect score: router + high form	0	0	#	#
Number and percent of children with chance score or below: router + low form	20	#	30	#

# Rounds to zero.

NOTE: The unweighted *n* is the rounded number of cases. Perfect scores are correct answers to all items administered, and chance scores are at the guessing level or below. Percentages are unweighted.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2012 and spring 2013; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.6.4 Score Statistics

Only IRT-based scores were produced for the science assessments. Table 5-33 presents summary statistics for the science theta, the *SEM* of theta, and the scale score, which indicate performance on the 64 unique items administered in the spring kindergarten, fall and spring first-grade, and fall and spring second-grade rounds.

Table 5-33. Science assessment statistics by IRT-based score, ECLS-K:2011 spring kindergarten, fall and spring first-grade, and fall and spring second-grade data collections: Spring 2011 and school years 2011–12 and 2012–13[15]

Variable	Description	*n*	Range of possible values	Weighted mean	Weighted standard deviation
X2STHETK2	X2 SCIENCE THETA	16,940	-8.0–+8.0	0.00	0.891
X3STHETK2	X3 SCIENCE THETA	5,180	-8.0–+8.0	0.43	0.932
X4STHETK2	X4 SCIENCE THETA	15,070	-8.0–+8.0	0.89	0.965
X5STHETK2	X5 SCIENCE THETA	4,720	-8.0–+8.0	1.18	0.954
X6STHETK2	X6 SCIENCE THETA	13,820	-8.0–+8.0	1.61	0.908
X2SSETHK2	X2 SCIENCE STD ERR OF THETA	16,940	0.0–6.0	0.71	0.081
X3SSETHK2	X3 SCIENCE STD ERR OF THETA	5,180	0.0–6.0	0.59	0.039
X4SSETHK2	X4 SCIENCE STD ERR OF THETA	15,070	0.0–6.0	0.59	0.032
X5SSETHK2	X5 SCIENCE STD ERR OF THETA	4,720	0.0–6.0	0.46	0.066
X6SSETHK2	X6 SCIENCE STD ERR OF THETA	13,820	0.0–6.0	0.46	0.050
X2SSCALK2	X2 SCIENCE IRT SCALE SCORE	16,940	0.0–64.0	28.07	7.526
X3SSCALK2	X3 SCIENCE IRT SCALE SCORE	5,180	0.0–64.0	31.90	8.653
X4SSCALK2	X4 SCIENCE IRT SCALE SCORE	15,070	0.0–64.0	36.29	9.198
X5SSCALK2	X5 SCIENCE IRT SCALE SCORE	4,720	0.0–64.0	39.32	8.782
X6SSCALK2	X6 SCIENCE IRT SCALE SCORE	13,820	0.0–64.0	43.35	8.366

NOTE: The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (1 for the fall kindergarten round, 2 for the spring kindergarten round, 3 for the fall first-grade round, 4 for the spring first-grade round, 5 for the fall second-grade round, and 6 for the spring second-grade round). X1 and X2 estimates weighted by W1C0; X3 estimates weighted by W3CF3P_30; X4 estimates weighted by W4CS4P_20; X5 estimates weighted by W6CF6P_2A0; and X6 estimates weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. 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) Restricted-Use Kindergarten–Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.6.5 Reliabilities

Table 5-34 presents reliability statistics for the spring kindergarten, fall and spring first-grade, and fall and spring second-grade science assessment scores (calculated in the same way as the reading reliability statistics, described in section 5.3.5). 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 reliabilities of the IRT-based scores from the science assessments are expected. Although the reliabilities for science are relatively lower than those in reading and mathematics, reliabilities of .8 and above are considered

[15] Plots of the IRT theta and scale score distributions are provided in appendix D.

acceptable based on the intended application of the ECLS-K:2011 test scores (Weiner, Schinka, and Velicer 2004).

Table 5-34. Science assessment reliabilities, ECLS-K:2011 spring kindergarten, fall and spring first-grade, and fall and spring second-grade data collections: Spring 2011 and school years 2011–12 and 2012–13

Score	Number of items	*n*	Weighted Reliability
Spring kindergarten			
IRT-based scores	20	16,940	.75
Fall first grade			
IRT-based scores	47	5,180	.83
Spring first grade			
IRT-based scores	47	15,070	.83
Fall second grade			
IRT-based scores	60	4,720	.83
Spring second grade			
IRT-based scores	60	13,820	.83

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 defined as the standard deviation of the posterior estimate) to total variance (across the sample). The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. 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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.7 Evaluating the Kindergarten Through Second-Grade Assessment Battery: Construct Validity and Measurement of Growth

This section addresses the issue of the validity of the national assessment scores as measures of child achievement and growth from kindergarten through second grade in the domains assessed. The validity issue is examined from several perspectives, each of which is discussed in detail below.

- ■ Is the difficulty of the tests suitable for children’s ability levels?
- ■ Do the tests measure the right content?
- ■ Do the data constitute a cohesive scale suitable for longitudinal measurement?

- What is the correlation of thetas from the same domain across rounds and grades (e.g., fall kindergarten reading with spring kindergarten reading or fall kindergarten mathematics with fall first-grade mathematics)? What is the correlation of thetas from different domains within a round (e.g., fall kindergarten reading with fall kindergarten mathematics)?

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

The two-stage adaptive tests were designed to maximize reliability within the available testing time by matching the test difficulty of the second-stage forms to children's ability level while minimizing the frustration or boredom that could occur if children received tests that were much too difficult, much too easy, or much too long. Evidence that both the two-stage and single-stage assessments contained items that were of appropriate difficulty for both the individual children taking them and, in the aggregate, for the rounds in which they were administered, can be found in analysis of the assessment data, specifically in the analysis of floor and ceiling effects (see sections 5.3 through 5.6). Assessments with minimal or no floor and ceiling effects are particularly important in a longitudinal study, where floor and ceiling effects can attenuate measurement of gain for the lowest and highest achieving children. No floor or ceiling effects were found for the assessments for any domain in any round; only a negligible number of children had below-chance or near-perfect scores on the two-stage assessments when considering performance on the routing and second-stage items combined or on the single-stage assessments. These psychometric results showing no significant floor or ceiling effects in any round indicate that the approach of administering the same assessment across rounds in the same year and the use of adaptive forms within a round was appropriate. The results also confirmed the successful selection of items of appropriate difficulty for the test takers.

5.7.2 Do the Tests Measure the Right Content?

Evidence for the appropriateness of the tests' content can be obtained from two sources: expert judgments and psychometric results. Chapter 2 includes a discussion of the development of the test frameworks, and chapter 4 includes a discussion of the design of the tests. Curriculum experts provided input with respect to the knowledge and skills that are both typically taught and developmentally important in kindergarten, first grade, and second grade. Test frameworks in each domain were developed to include the knowledge and skills recommended by the curriculum experts, and test items in each assessment were selected to conform as closely as possible to framework specifications. The field test item pools and the sets of items selected for the national assessments were reviewed by experts, and the content and presentation of items were modified in response to their recommendations.

A psychometric analysis of the appropriateness of test content included a review of the common functionality of each item and floor and ceiling effects across rounds. The common functioning of items administered in multiple rounds of data collection indicate that although on average, the ability levels of the children increase with each subsequent round of data collection, the relative difficulties of the items do not. This stability of the item difficulty estimates across rounds provides further evidence that the items are administered in the appropriate rounds. As described in section 3.5, IRT calibration allowed for the estimation of performance on each item for *both* rounds, fall and spring, of the kindergarten, first-grade, and second-grade years. The match of assessment forms to estimated performance (as evidenced by the common functioning of items, and the lack of significant floor or ceiling effects in any round in any domain) suggests that the content of the tests reflected what children had been learning during the school year, and their progression of skills longitudinally. If the assessments included content that was too easy at the time of the data collections, a ceiling effect would likely be observed, since many children would respond correctly to most of the items. Similarly, if the assessments included content that was too difficult at the time of the data collections, many children would respond incorrectly to most of the items, resulting in a floor effect.

5.7.3 Do the Data Constitute a Cohesive Scale Suitable for Longitudinal Measurement?

Whether the data collected in each round of administration were suitable for longitudinal measurement was explored in three ways by examination of (1) IRT *a* parameters, (2) differential item functioning analysis of common items, and (3) increase in proportion correct over time.

Examination of IRT *a* parameter estimates is one way to examine whether the items constitute a cohesive, valid measure of the assessment domain. IRT *a* parameters and the item-test biserial correlation are approximately monotonic increasing functions of each other (Lord 1980). Thus, higher *a* parameters are related to higher item-test correlations and, therefore, the *a* parameter estimates indicate how strongly each item is related to the underlying construct being measured by the test. (See section 3.2 for a description of the IRT *a* parameter, and appendix C for a table of the item parameter estimates for each domain.) While this section presents information on the *a* parameters, factor analysis can also be used to evaluate the cohesiveness of the assessment. The factor analysis results are reported in the context of the discussion on unidimensionality, in section 5.2.1.

Examination of the *a* parameter is warranted because of the two-stage design of the assessment, where not every test taker answered *all* of the assessment items. If each test taker had answered *all* of the assessment items on all forms for a given domain in all rounds of data collection, it would be possible to measure the cohesiveness of the scale by reviewing alpha coefficients and item biserials. However, because of time constraints and the age of the children in the study, it would have been neither reasonable nor practical to administer every item to every child in every round. The IRT *a* parameter estimates provide the same type of insight into the cohesiveness of a set of test items as do alpha coefficients and item biserials. As discussed in section 3.2, this parameter represents item discrimination, or the ability of an item to discriminate between children whose ability level is above or below the calibrated difficulty of the item. Values above 1.0 for most of the items in a test constitute evidence that there is a strong underlying factor being measured by the test. However, the presence of some items with IRT *a* parameters less than 1.0 does not necessarily indicate non-unidimensionality. Items that are too easy or too difficult for the majority of the sample may have *a* parameters lower than 1.0, but still contribute to a single, underlying factor.

Across the kindergarten, first-grade, and second-grade assessments, there was a total of 71 unique items contributing to the scoring model for the reading assessment. Of the 71 calibrated items in the reading scale (17 polytomous and 54 dichotomous), 47 have *a* parameter estimates greater than 1.0. Items with *a* parameter estimates slightly below 1.0 are related to letter-sounds, addition of phonemes, blending, segmentation, rhymes, conventions of print, and three of the more challenging passage sets. Those with the lowest *a* parameter estimates are sound-matching items and one challenging passage set. Vocabulary items showed high and low *a* parameter estimates, depending on the difficulty of the item (i.e., the more challenging items had lower *a* parameter estimates while the easier vocabulary words had higher *a* parameter estimates). Nearly all of the items tapping letter recognition, decoding, and about half of the passage sets had *a* parameter estimates above 1.0.

Results for mathematics were similar. Across the kindergarten, first-grade, and second-grade assessments, there was a total of 97 unique items contributing to the scoring model for the mathematics assessment, with 62 of the 97 items having *a* parameter estimates above 1.0. Items with *a* parameter estimates slightly below 1.0 were varied in type. Items having the lowest *a* parameter estimates were generally either fairly easy or fairly challenging for the sampled children.

In science, across the kindergarten, first-grade, and second-grade assessments, there was a total of 60 unique items contributing to the scoring model for the science assessment. In the science assessment, 46 of the 60 items had *a* parameter estimates above 1.0. Fourteen of the 60 items administered had *a* parameter estimates lower than 1.0. The diverse content assessed, combined with the lower variability in children's ability levels in science, resulted in relatively more items with *a* parameter estimates that were lower than ideal.

In the SERS assessment (administered in kindergarten and first grade only), 13 of the 31 items administered had *a* parameter estimates greater than 1.0. The small subsample of test takers and the low variation in the ability levels of those test takers (as evidenced in the standard deviations of theta) resulted in more than half of the items with *a* parameter estimates that were less than ideal.

The proportion of items with lower *a* parameter estimates in the SERS seems somewhat contradictory to the results from the component analyses (section 5.2.1), where the SERS was shown to be highly unidimensional. The kindergarten SERS data showed a primary factor accounting for 30.4 percent of the variance and that 22 items loaded to that primary factor, suggesting one strong underlying factor. Additionally, the number of items contributing to the scaling of the assessment and the range of theta for the SERS assessment was smaller than those in the reading and mathematics assessments. The smaller ranges in the scales, and thus lower possible variance, resulted in relatively more items with lower *a* parameter estimates in the SERS assessment.

In addition to examining the *a* parameters in support of one underlying factor, differential item functioning (DIF) contrasting data from the fall and spring at each grade level indicate that the items are common-functioning longitudinally across rounds within a given grade and that the IRT model appropriately represents the test data collected in each round.

Furthermore, evidence that the IRT models produce scores that can be validly used to model growth is shown in the increase in proportion correct over time (as illustrated across rounds from the classical item analysis proportion correct calculations), and that the increases that took place are consistent with the model (given the varied content and difficulties of the items).

5.7.3.1 Precision of Gain Scores Used to Model Growth

As discussed in section 5.2.4.1, the estimated standard error of theta provides a measure of uncertainty of the theta score estimate for each child. Unlike classical item theory, which assumes the precision of the scores is usually consistent across all examinees, IRT procedures provide an estimate of the accuracy of the theta estimate for each test taker.

The standard error of theta of the gain score is a function of the standard errors of the thetas used to compute the gain score, and is defined as the square root of the sum of the squared standard errors of theta for each score:

$$S_{G(\theta_1, \theta_2)} = \sqrt{S_{\theta_1}^{2} + S_{\theta_2}^{2}} \ ,$$

where

S = standard error of measurement;
G = gain score;
θ_1 = ability of the test taker at point 1; and
θ_2 = ability of the test taker at point 2.

The standard error for the gain score is thus larger than the standard error of theta from either of the two component thetas (assuming these are both nonzero). Gain scores are less precise than the scores that they are derived from, and the standard errors of the gain scores will vary, depending upon the theta values at both points in time. Adding and subtracting twice the standard error estimate from the theta gain score estimates provides an approximate 95 percent confidence interval or range of values that is likely to include the child's true theta gain.

5.7.4 Relationship of the Cognitive Test Scores to Scores in Different Rounds and Different Subjects

Table 5-35 shows the correlations of scores for assessments in the same subject across rounds. Construct validity of the assessments can be demonstrated by these correlations. Correlations between more similar measures in the same domain should be high, while correlations between less similar measures in the same domain should be low. For example, within the same domain over time, correlations between assessment uses in adjacent time points should be higher than correlations between assessments used at time points further apart because adjacent assessments have a higher proportion of common items. As can be seen in table 5-35, correlations are generally highest near the diagonal and get progressively lower toward the lower left corner of each set. In other words, scores in each subject appear to be most closely related to the most recent or subsequent score, and less closely related to rounds that are more distant. There is one notable deviation from this pattern, however. The SERS correlations do not necessarily follow the expected trend of correlations being higher at adjacent time points, but this is likely due to the very low numbers of observations, and thus the variance, in first-grade administrations.

Correlations of scores *across* subjects *within* rounds are presented in table 5-36. It could be expected that across domains within the same time point, scores on cognitive assessments such as these that measure knowledge and skills in theoretically related domains and that reflect an underlying level of ability would be related. The correlations are consistent with those of the ECLS-K and ECLS-B studies assessing students in reading, mathematics, and/or science.

Table 5-35. Correlations of IRT theta score across rounds, by subject, ECLS-K:2011: School years 2010–11, 2011–12, and 2012–13

Subject	Fall kinder-garten	Spring kindergarten	Fall first grade	Spring first grade	Fall second grade	Spring second grade
Reading						
Fall kindergarten	1.00					
Spring kindergarten	.79	1.00				
Fall first grade	.76	.85	1.00			
Spring first grade	.70	.81	.87	1.00		
Fall second grade	.67	.76	.83	.90	1.00	
Spring second grade	.65	.74	.79	.87	.90	1.00
SERS						
Fall kindergarten	1.00					
Spring kindergarten	.69	1.00				
Fall first grade	.60	.76	1.00			
Spring first grade	.74	.79	.85	1.00		
Fall second grade	†	†	†	†	†	†
Spring second grade	†	†	†	†	†	†
Mathematics						
Fall kindergarten	1.00					
Spring kindergarten	.81	1.00				
Fall first grade	.77	.84	1.00			
Spring first grade	.76	.82	.86	1.00		
Fall second grade	.73	.80	.83	.88	1.00	
Spring second grade	.72	.79	.80	.86	.90	1.00
Science						
Spring kindergarten	†	1.00				
Fall first grade	†	.80	1.00			
Spring first grade	†	.77	.85	1.00		
Fall second grade	†	.74	.80	.84	1.00	
Spring second grade	†	.72	.78	.82	.87	1.00

† Not applicable.

NOTE: All estimates unweighted. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. IRT = item response theory. 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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten--Second Grade Data File and Electronic Codebook NCES 2015-050).

Table 5-36. Correlations of IRT theta score across subjects, by round, ECLS-K:2011: School years 2010–11, 2011–12, and 2012–13

Round	Reading and mathematics	Reading and science	Mathematics and Science
Fall kindergarten	.78	†	†
Spring kindergarten	.76	.55	.60
Fall first grade	.76	.59	.65
Spring first grade	.75	.64	.68
Fall second grade	.73	.66	.69
Spring second grade	.73	.70	.72

† Not applicable.
NOTE: All estimates unweighted. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. 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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten--Second Grade Data File and Electronic Codebook (NCES 2015-050).

5.8 Selection and Use of Scores

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

5.8.1 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. Theta scores from all rounds are on the same metric. Therefore, an analyst looking at growth from fall kindergarten to fall first grade could subtract the fall kindergarten score from the fall first grade 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.[16] 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. Scale scores from all rounds are on the same metric. Therefore, an analyst looking at growth from fall kindergarten to fall first grade could subtract the fall kindergarten score from the fall first grade 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 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 in overall performance on the reading assessment, regardless of a child's home language.

[16] 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.

- 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.

5.8.2 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 collections, and from January to July for the spring rounds. Children assessed later in a data collection period in a particular grade level, for example in December during a 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). Two measures of executive function were included in the kindergarten, first-grade, and second-grade direct child assessment battery: the *Dimensional Change Card Sort (DCCS)* (Zelazo 2006, Zelazo et al. 2013), assessing children's cognitive flexibility, and the Numbers Reversed subtest of the *Woodcock-Johnson III Tests of Cognitive Abilities* (Woodcock, McGrew, and Mather 2001), assessing working memory. The same versions of the *DCCS* and the Numbers Reversed tasks were administered in fall and spring of the kindergarten year and fall and spring of first grade. Because the physical version of the *DCCS* used in kindergarten and first grade was a task designed for younger children, the *DCCS* was changed in the fall of second grade to a computerized version of the task that was appropriate for measuring cognitive flexibility for older children. The version of the Numbers Reversed task remained the same as the version used in the early data collection rounds.

6.1 *Dimensional Change Card Sort* (*DCCS*)

The *DCCS* is used to collect information on children's cognitive flexibility. In the kindergarten and first-grade data collections, the *DCCS* was administered as a physical, table-top card sort with the items administered by a trained assessor. In the second-grade data collections, a computerized version of the *DCCS* developed for the National Institutes of Health Toolbox for the Assessment of Neurological and Behavioral Function (NIH Toolbox) was administered. As noted above, this task was administered differently in the second-grade data collections than it was in the earlier collections to remain age-appropriate. This section describes both administration methods and the types of data generated by each method.

In the version of this task used in kindergarten and first grade, children were asked to sort a series of 22 picture cards according to different rules. Each card had a picture of either a red rabbit or a blue boat. The children were asked to sort each card into one of two trays depending on the sorting rule they had been told to use. One tray had a picture of a red boat and the other had 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 was 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. The Color Game trials are referred to as the "pre-switch" trials because they occur first, before the child is asked to change or "switch" sorting rules. In the second game, the Shape Game, the child was asked to change sorting rules and 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. The Shape Game trials are referred to as the "post-switch" trials because they occur after the change or "switch" from sorting by color to sorting by shape. If the child correctly sorted four of the six cards in the Shape Game, then he or she moved on to the third game: the Border Game. In the Border Game, the sorting rule (by color or by shape) depended on whether the card had a black border around the edges. If the card had a border, the child was to sort by color; if there was no border on the card, the child was to sort by shape.

Kindergarten and first-grade item-level data for this version of the *DCCS* are provided in the ECLS-K:2011 kindergarten—second grade data file. There are six variables with item-level data for the Color Game, six variables with item-level data for the Shape Game, and six variables with item-level data for the Border Game. There were four practice items administered to children, but the item-level data from these practice items are not included in the data file. The item-level data for the Color, Shape, and Border Games are scored "correct" (i.e., the card was sorted into the correct tray according to the sorting rule) or "incorrect" (i.e., the card was 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 in either the Color or Shape Game to advance to this set of items in the assessment. The "not administered" code is different than a system missing code in that only those children who were administered the *DCCS* can have a "not administered" code. If a child was not administered the *DCCS* at all, his or her data for the *DCCS* scores would be coded as missing. 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." Similarly, variable names for the item-level data from the fall and spring first-grade assessments begin with "C3" and "C4," respectively.

Using scoring rules provided by the developers, four scores were developed from the *DCCS* data for the fall and spring kindergarten and the fall and spring first-grade rounds of data collection: the pre-switch score, the post-switch score, the Border Game score, and a total score (identified as an overall score on the data file). The pre-switch score is the number of cards the child correctly sorted by color during the first phase of the assessment. The post-switch score is the number of cards the child correctly sorted by shape after switching from sorting by color to sorting by shape. 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.[1] A final combined scale score reflects the total accuracy for the three tasks (i.e., the total number of cards sorted correctly in the Color, Shape, and Border Games), which results in a maximum score of 18 correct. The developer of the *DCCS* recommends using the total accuracy score to assess performance.

The total *DCCS* scores for kindergarten and first grade included in the kindergarten–second grade data file are calculated differently than what was recommended for calculation of the total *DCCS* score in the base-year User's Manual. The *User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic Codebook, Public Version* (NCES 2015-074) (Tourangeau et al. 2015) indicated that researchers could create a single *DCCS* composite score by summing the post-switch score and the Border Game score and use that combined score in analyses. This composite score does not include information from the pre-switch score. After the release of the kindergarten-year file, further consideration was given to the fall and spring kindergarten data, where 1,038 cases scored 0 on the post-switch score in the fall of kindergarten and 457 cases scored 0 on the post-switch score in the spring of kindergarten. Following consultation with the *DCCS* developer, it was subsequently decided to include the pre-switch scores in the calculation of the total *DCCS* score in order to better capture variability at the lower ability levels. Therefore, the total *DCCS* scores for kindergarten and first grade (X*DCCSTOT)[2] included in the kindergarten–first grade and kindergarten–second grade data files reflect children's performance across all 18 trials (i.e., the pre-switch or Color Game trials, the post-switch or Shape Game trials, and the Border Game trials).[3]

In kindergarten and first grade, the *DCCS* was administered in Spanish for children routed through the Spanish assessment. Data from the English and Spanish administrations are combined into the same item-level variables and into the same score variables. Researchers who want to account for language of administration in their analyses can use the data flag provided on the data file for each round (X*FLSCRN) to identify which cases were administered the *DCCS* in English and which cases were administered it in Spanish. All children were administered the assessments in English in the second-grade

[1] All children initially attempted six Color Game trials, and then moved to the Shape Game. 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 higher 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 Zelazo 2006.

[2] In variable names, the "*" is to be substituted with the appropriate data collection round number. In this case, the variables for the fall and spring kindergarten (rounds 1 and 2) and the fall and spring of first grade (rounds 3 and 4) are X1DCCSTOT, X2DCCSTOT, X3DCCSTOT, and X4DCCSTOT.

[3] The total *DCCS* scores for kindergarten and first grade that are included on the kindergarten–first grade data file are calculated in the same way as the total scores for kindergarten and first grade on the kindergarten–second grade file. They are the total correct across all 18 trials, namely the pre-switch, post-switch, and Border Game trials.

collections. Therefore, the second-grade *DCCS* scores for all children are based on an English administration of the assessment, and data flags to indicate language of administration in second grade (rounds 5 and 6) are not provided in the data file.[4]

Because the version of the *DCCS* used in kindergarten and first grade would have been too easy for the majority of the study children during the second-grade rounds, children were administered a new, age-appropriate, computerized version of the *DCCS* in which the "cards" are presented on a computer screen and children sort them into virtual "piles" on the screen using keys on the keyboard to indicate where to place each card. The ECLS-K:2011 Executive Function Content Review Panel[5] reviewed these plans and supported transitioning to a computerized version of the *DCCS*. The Content Review Panel members indicated that the results of the tabletop and computerized versions could be compared across rounds and suggested that standardized scores would facilitate the comparisons.

The computerized task was developed as part of the National Institutes of Health Toolbox for the Assessment of Neurological and Behavioral Function (NIH Toolbox) and is appropriate for ages 3 through 85[6] (Zelazo et al. 2013). The task had been under development and became available in time to be incorporated into the second-grade data collections. The NIH Toolbox *Dimensional Change Card Sort* Test (NIH Toolbox *DCCS*) consists of 40 trials, including 5 pre-switch trials (where are children are asked to sort by one dimension, e.g., color), 5 post-switch trials (where children are asked to sort by a different dimension, e.g., shape), and 30 mixed-block trials (in which the sorting dimension, either color or shape, varies by trial). The pre-switch and post-switch trials on the NIH Toolbox *DCCS* are based upon the same pre-switch and post-switch trials in the physical version of the *DCCS* that the ECLS-K:2011 administered in kindergarten and first grade.[7] However, unlike the physical version administered in ECLS-K:2011, which has the Border Game following the post-switch trials to increase the level of difficulty, the NIH Toolbox *DCCS* administered a set of mixed-block trials to increase the difficulty of the task to make it appropriate for a much broader age range.

The NIH Toolbox *DCCS* has two different start points based on the age of the child in order to limit administration time. Testing conducted in the development of the NIH Toolbox *DCCS* indicated

[4] More information about how children's home language affected children's routing through the assessment battery in each round of data collection is provided in chapter 5.

[5] The Executive Function Content Review Panel was a group of experts who were asked to advise study design staff on the best ways to measure executive function for younger children. The group provided input on the measures for second-grade in the spring of 2011.

[6] See www.nihtoolbox.org for additional information about the NIH Toolbox *DCCS* and about the NIH Toolbox for the Assessment of Neurological and Behavioral Function.

[7] There were 6 pre-switch and 6 post-switch trials in the physical version of the *DCCS* administered in kindergarten and first grade. The NIH Toolbox *DCCS* includes 5 pre-switch and 5 post-switch trials.

that 8-year-olds typically scored at ceiling on the pre-switch and post-switch trials. Consequently, children under age 8 begin with the pre-switch trials, and children age 8 and older begin with the mixed-block trials and are given credit in the scoring for completing the pre-switch and post-switch trials accurately. It is important to note that all ECLS-K:2011 children in second grade were administered the version of the NIH Toolbox *DCCS* for ages 8 years and older, regardless of their age at the time of assessment. This was done so that all study children would receive the same version of the *DCCS* task in the fall and spring of second grade and in later rounds of data collection. This has some implications for the data collected, which are discussed below in section 6.1.2.

The construct assessed in the physical version and the computerized version of the *DCCS* is the same–cognitive flexibility. However, the way the construct is assessed and the scoring differ across the versions. One key difference between the two versions is that the computerized version captures data on the amount of time in milliseconds (msec) that it takes the child to complete any given item; it is not possible to accurately measure reaction time at the necessary level of precision in the physical version. Therefore, the computerized version supports the use of both accuracy of sorting and reaction time to assess overall performance while the physical card sort assesses performance by accuracy alone.

Using accuracy alone to measure cognitive flexibility is developmentally appropriate for younger ages. However, as children get older, scoring that takes both accuracy and reaction time into consideration becomes necessary to remain developmentally appropriate. The well-studied developmental phenomenon known as the speed-accuracy tradeoff results in different but comparable ways of measuring the “cost” of executive-function demands that are appropriate at different developmental stages. (See Zelazo 2013.) Younger children tend to respond rapidly with lower accuracy; because there is sufficient variability in performance based on accuracy alone, the physical card sort is generally appropriate for assessing cognitive flexibility in younger children. In contrast, older children and adults typically learn to slow down while sorting the cards in order to avoid making mistakes, which results in their sorting with high accuracy. It becomes important to factor in reaction time once the child begins to master the speed-accuracy tradeoff and is able to slow responding enough to maintain high accuracy. When an individual is able to maintain high accuracy on the task, then the speed at which the individual responds becomes a better indicator of executive function. Sorting accurately more quickly is an indicator of higher cognitive flexibility once relatively high accuracy is achieved. Thus, using accuracy alone and using accuracy in combination with reaction time are comparable indices of executive function, if considered developmentally.

In each of the 30 mixed-block trials administered to children in the ECLS-K:2011, the children were presented with a stimulus picture of a ball or truck that was either yellow or blue. A prerecorded female voice announced the sorting rule to be used for that trial ("color" or "shape") as the appropriate word "color" or "shape" was briefly displayed in the center of screen. Next, the stimulus picture was displayed in the center of screen where the word had just appeared. Children then selected one of two pictures at the bottom of the screen (a blue ball on the left or a yellow truck on the right) that was either the same shape or the same color as the stimulus picture, depending on whether the shape or color sorting rule was in effect for the trial. Children indicated their choice of picture by pressing the arrow key on the laptop keyboard that was associated with the picture; the left arrow key was used to select the picture on the left side of the screen and right arrow key was used to select the picture on the right side of the screen. Children were instructed to use just one pointer finger to press the arrow keys. They were asked to return their pointer finger to the button in between the left and right arrow keys (marked with a fuzzy sticker, and so identified as the "fuzzy button") in between trials to standardize the start location for every child's finger, with the goal of maximizing accuracy in the measurement of response time.[8] Both reaction time to sort the card and accuracy of its placement according to the sorting rule in effect for the trial were recorded by the computer program.

The sorting rules (i.e., to either sort by shape or color) were intermixed across the trials, and one rule was more common than the other. The shape rule was used for 23 trials while the color rule was used in 7 trials. For example, the child may be asked to sort by shape for 4 trials in a row, then to sort by color on trial 5, and then to sort by shape on trials 6 and 7. One sorting rule was presented more frequently in order to build a response tendency (i.e., a response that is "preferred" because it happens more frequently, resulting in a predisposition to respond in that manner). Depending upon how well a child can slow down in order to maintain accuracy, a predisposition to sort by the dominant rule (i.e., shape) can result in either more errors or a slower reaction or response time on nondominant trials because it is necessary to inhibit the dominant response (i.e., sorting by shape) in order to shift to the less frequent sorting rule (i.e., color). The "cost" associated with the shift from a more frequent rule (the "dominant" rule) to a less frequent rule (the "nondominant" rule) tends to differ by the age of the participant (Davidson et al. 2006). The "cost" to younger children is that they tend to make more errors on the nondominant rule trials; that is, they do not demonstrate the cognitive flexibility to make the switch in rule even when prompted. Younger children do not tend to slow themselves down in favor of higher accuracy, and, therefore, accuracy is a better metric of performance for young children (Zelazo et al.

[8] Experts from the ECLS-K:2011 Content Review Panel who had experience with this and/or similar tasks for children in the age range of the ECLS-K:2011 study children recommended that a "fuzzy button" be used for this study. The NIH Toolbox *DCCS* does not use a "fuzzy button."

2013). In contrast, older children and adults tend to demonstrate a speed-accuracy tradeoff; they slow down the pace at which they respond in order to maintain accuracy. Thus, the "cost" to older children and adults is seen in reaction time on the nondominant rule trials. The formula used to produce scores from the data collected by the computerized *DCCS* factors in reaction time on the infrequent or nondominant trials when a child demonstrates sufficiently accurate performance across all the test trials, defined as being accurate on more than 80 percent or more of the trials (Zelazo et al. 2013). Thus, the computerized DCCS provides a measure of performance through this developmental shift to learning to trade speed for accuracy. More information on scoring is provided below.

The 30 test trials were only administered to children who successfully completed the practice portion of the *DCCS*. The practice consisted of a minimum of 8 trials and a maximum of 24 trials, depending upon how quickly the child demonstrated that he or she understood the task. For the first set of practice trials, the assessor instructed the child how to sort by shape using text automatically presented on the *DCCS* screen that was read by the assessor along with additional standardized instructions presented by the assessor that were not on the computer screen. Following the instructions, the computer administered four practice trials asking the child to sort by shape. If the child sorted at least three of the four items correctly by shape, he or she progressed to the color practice. If the child sorted more than one item in the set of four incorrectly, he or she was presented with a second set of four practice items. If the child failed to sort three of four items correctly by shape in the second set of practice items, he or she was presented a third set; failure of this third set ended the *DCCS* program before any actual scored trials were presented.

Once a child passed the shape practice trials, the assessor instructed on how to sort by color, and the computer presented 4 to 12 practice trials asking to sort by color. Like the shape practice trials, up to three sets of four items could be presented before the *DCCS* advanced to the scored trials. If the child was not able to pass the color practice, the *DCCS* program ended after the third set of color practice items, again before any actual scored trials were presented.

In contrast with the scored trials, the practice trials maintained one sorting rule for all items presented in succession until practice for the rule was complete. An additional difference between the practice and scored trials was that the stimulus pictures in the practice trials were white or brown rabbits and boats, while the stimulus pictures were yellow or blue trucks or balls in the scored trials.

Item-level data for the 30 scored trials are included in the data file. They are provided in three blocks of 30 items for each participant that indicate: (1) correct versus incorrect responses; (2) the type of trial, reported as dominant (most frequently presented but not included in reaction time scores; shape is the dominant sorting rule) or nondominant (less frequently presented and used to calculate reaction time scores; color is the nondominant sorting rule); and (3) reaction times reported in milliseconds. Variable names for the item-level data begin with “C5” for fall second grade and “C6” for spring second grade.

The total computed score reported for the second-grade *DCCS* is derived using a formula provided by the task developer and follows the scoring algorithm used for this task in the NIH Toolbox (see the *NIH Toolbox Scoring and Interpretation Guide*, Slotkin, Nowinski et al. 2012, for additional information on scoring). This score is a function of both accuracy and reaction time to take into consideration speed-accuracy tradeoff, which was discussed above. Scores range from 0 to 10, with accuracy (0 to 5 units) and reaction time (0 to 5 units) taken into account in the computation of the scores. Accuracy is considered first. If the child’s accuracy rate is less than or equal to 80 percent, the child’s total computed score is based entirely on accuracy. If the child’s accuracy rate is more than 80 percent, the child’s total computed score is based on a combination of accuracy and reaction time.

The accuracy score factored into the computation of the total score can range from 0 to 5. There are a total of 40 possible accuracy points that are scaled down to a maximum score of 5: for each correct response, the child earns a score of .125 (5 points divided by 40 trials). Because all children used the start point of the *DCCS* for children 8 years and older, each child was administered the 30 mixed-block trials, and each child who successfully passed the practice items was automatically given 10 accuracy points for the 5 pre-switch and the 5 post-switch trials of the *DCCS* that were not administered. Therefore, the accuracy component of the total computed *DCCS* score is calculated as follows:

DCCS Accuracy Score = 0.125 * Number of Correct Responses[9]

If the child’s accuracy rate is higher than 80 percent, a reaction time score is added to the child’s accuracy score.[10] Like the accuracy score, the reaction time score ranges from 0 to 5 points. The

[9] The number of correct responses = 10 + the number of correct trials out of the 30 mixed block trials. Because 10 accuracy points are automatically awarded due to the chosen start point for the task, it is not possible for ECLS-K:2011 children to get an accuracy score of 0. Therefore, the minimum possible value for the *DCCS* accuracy score is 1.25, and the maximum possible *DCCS* accuracy score is 5.

[10] The criterion of *greater than* 80 percent accuracy is calculated based on all 40 trials (30 administered trials plus the 10 trials not administered). That is, 80 percent of 40 trials is 32 items. However, this can also be thought of in terms of how many items out of the 30 administered trials are required. If the criterion is 80 percent of the 40 trials, this translates to 23 of the 30 administered trials. For example, if a child responds accurately on 23 of the 30 mixed block trials, the child’s accuracy rate equals 82.5 percent (10 points automatically awarded for the pre-switch and post-

reaction time component of the total computed score for the computerized *DCCS* is computed using the child's median reaction time to correct nondominant trials (i.e., the trials with the less frequently used sorting rule, color), following the same scoring algorithm outlined in the scoring manual for the NIH Toolbox *DCCS* (Slotkin, Nowinski et al. 2012). First, the median reaction time is calculated based on reaction times for correct nondominant trials with reaction times greater than or equal to 100 msec and within plus or minus three standard deviations from the child's mean reaction time on the correct nondominant trials. The minimum median reaction time allowed is 500 msec; the maximum median reaction time is 3,000 msec. If the child's median reaction time falls outside this range, the child's median reaction is set to the minimum or maximum allowable range: reaction times between 100 msec and 500 msec are set to 500 msec and reaction times between 3,000 msec and 10,000 msec (the maximum trial duration) are set to 3,000 msec. A log (base 10) transformation is applied to the median reaction times to create a more normal distribution.

The log values are then algebraically rescaled to a 0 to 5 range and then reversed such that faster (better) reaction times have higher values and slower reaction times have lower values. The formula for rescaling the median reaction times is the following:

$$\text{Reaction time score} = 5 - \left(5 * \left[\frac{\log RT - \log(500)}{\log(3000) - \log(500)}\right]\right)$$

where *RT* is the median reaction time on nondominant trials within set outer limits.[11]

To summarize, the total computed score on the computerized *DCCS* (identified as an overall score on the data file) is equal to the child's accuracy score if the child's accuracy rate is less than or equal to 80 percent. If the child's accuracy rate is greater than 80 percent, the child's total computed score is equal to the child's accuracy score plus the child's reaction time score, which is derived from the child's reaction time on correct nondominant trials as described above. Additional details on the calculation of the computed score are available in the *NIH Toolbox Scoring and Interpretation Guide* (Slotkin, Nowinski et al. 2012) and the *NIH Toolbox Technical Manual* (Slotkin, Kallen et al. 2012).

switch trials plus the 23 correct mixed block trials divided by 40; 33/40 = .825). In this example, the child's accuracy score would be [(10 + 23) * .125] = 4.125. Because the accuracy rate is *greater than* 80 percent, the child's reaction time score would be added to this accuracy score to obtain the overall computed score for the *DCCS*. Alternatively, if the child responded accurately on 22 of the 30 mixed-block trials, the child's accuracy rate would equal 80 percent and, therefore, the child's accuracy is not *greater than* 80 percent and the child's overall score would be based solely on accuracy (overall computed score = [(10 + 22) * .125] = 4).

[11] The median reaction time (*RT*) used to calculate the reaction time score falls within the range of 500 msec through 3,000 msec. Calculation of the median score requires a minimum of at least one correct nondominant trial reaction time that is greater than 100 msec. When the child reached the accuracy threshold for including the reaction time component in the scoring but did not have any within-range reaction times on correct nondominant trials, the child's overall computed score on the *DCCS* was set equal to the child's accuracy score, and the reaction time was not factored into the child's score.

The overall computed score on the computerized *DCCS* cannot be compared with the kindergarten and first-grade total scores (X1DCCSTOT, X2DCCSTOT, X3DCCSTOT, and X4DCCSTOT) without first standardizing the scores. The kindergarten and first-grade total scores indicate the number of items a child sorted accurately and have a potential range of 0 to 18. The second-grade computed scores (X5DCCSSCR and X6DCCSSCR) range from 0 to 10, with weight given to accuracy (0 to 5 units) and reaction time (0 to 5 units) in the computation of the scores. The total computed scores for the physical card sort can be used to examine change across rounds that use the physical cards (e.g., performance in the fall of kindergarten can be directly compared to performance in the spring of first grade). The total computed scores for the computerized *DCCS* can be used to examine change across rounds that use the computerized *DCCS* (i.e., performance in the fall of second grade can be directly compared to performance in the spring of second grade).

It is important for researchers using the DCCS data to be aware of the characteristics of the overall DCCS scores and determine how best to use these scores in their analyses. As noted above, the NIH-developed scoring model computes scores differently depending on sorting accuracy. The use of this scoring model with the data collected from children in the ECLS-K:2011 resulted in a non-normal distribution, with approximately 14 percent of children in the fall second-grade data collection and 10 percent of children in the spring second-grade data collection who have a computed overall score failing to achieve greater than 80 percent accuracy and, therefore, having their score calculated based solely on accuracy. The remaining children have scores calculated based on both accuracy and reaction time. The non-normal distribution may be problematic for statistical analyses. For this reason, users may want to run analyses that do not use the overall score as is with the full sample. For example, users could conduct their analyses separately for the two groups of children so that each analysis only includes children with scores calculated in the same way, or they may decide to limit their analyses to only one group. Another option is for users to analyze all children using the score indicating accuracy alone, recognizing that this score is highly skewed, as most children were able to sort the cards with at least 80 percent accuracy. Users may also want to consider investigating alternative scoring models using the item-level accuracy and reaction time data available on the data file. The decision about how best to use the DCCS overall score in analysis is left to the user, given the research questions being addressed. Analysts may choose to examine other ways researchers have analyzed data with similar distributions, or other executive function or card sort data, in deciding how best to utilize the ECLS-K:2011 DCCS data.

For each round, there is a flag variable on the data file that indicates the presence of a valid *DCCS* score (X*DCCSFLG). For the physical card sort in the kindergarten and first-grade rounds of data collection, these flags indicate the presence of data on the *DCCS* combined score (X*DCCSTOT). The use of computers for the administration of the *DCCS* in second grade allowed the completion flags (X5DCCSFLG and X6DCCSFLG) to provide additional detail that was not available for kindergarten and first grade. The values indicate whether the task was administered, whether the total computed *DCCS* score is present, and, if a score is not present, the reason why it not present. Reasons why a score is not present when the *DCCS* was administered include failing the shape practice trials, failing the color practice trials, and having an administrative breakoff (the assessor ended the task) either before or after passing the practice trials. Administrative breakoffs could have occurred for a variety of reasons such as an external event (fire drill, child needed to return to class) that interrupted a test session. Note that the Shape Game preceded the Color Game during the practice trials. Table 6-1 presents a description of the completion codes for the physical card sort (kindergarten and first grades), along with the percentage of cases with data on the *DCCS* combined score. Table 6-2 presents a description of the completion codes for the computerized *DCCS* administered in the fall and spring of second grade, along with the percentage of cases that are assigned each code.

Table 6-1. Percentage of round participants assigned each code on the kindergarten and first-grade completion flag variables for the physical version of the *Dimensional Change Card Sort*: School years 2010–11 and 2011–12

X*DCCSFLG	Value	Fall kindergarten (X1DCCSFLG)	Spring kindergarten (X2DCCSFLG)	Fall first grade (X3DCCSFLG)	Spring first grade (X4DCCSFLG)
False	0	1.14	.38	.15	.15
True	1	98.86	99.62	99.85	99.85

NOTE: Detail 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, spring 2011, fall 2011, and spring 2012.

Table 6-2. Percentage of round participants assigned each code on the completion flag variables for the computerized version of the *Dimensional Change Card Sort* for fall and spring second grade: School year 2012–13

X*DCCSFLG	Value	Fall second grade (X5DCCSFLG)	Spring second grade (X6DCCSFLG)
Not administered	0	.36	.24
DCCS computed (overall) score present	1	99.43	99.45
Failed Shape Game practice	2	.13	.15
Failed Color Game practice	3	.08	.06
Breakoff before passing practice trials	4	#	.04
Breakoff after passing practice trials	5	#	.06

# Rounds to zero.
NOTE: Detail 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 2012 and spring 2013.

6.1.1 Mean Scores

This section presents the overall means for the *DCCS* scores in first and second grade. For the first-grade rounds of data collection, there are four scores for the *DCCS* for each round: the pre-switch score, the post-switch score, the Border Game score, and a combined/total score. For the second-grade rounds, there are three scores for each round: the total computed score which is a function of accuracy and reaction time and the two component scores that are summed to create the total score, the accuracy component score and the nondominant reaction time component score. (See section 6.1 above for detailed information on how these scores are derived.)

Table 6-3 presents the variable names, descriptions, value ranges, weighted means, and standard deviations for the pre-switch, post-switch, Border Game, and combined/total scores for the fall first-grade and spring first-grade *DCCS* as well as for kindergarten. Means for the kindergarten scores are presented here because the combined/total scores were not included on the kindergarten data file and, therefore, they were not included in the *ECLS-K:2011 Kindergarten Psychometric Report* (Najarian et al. 2018). The kindergarten scores for the Color Game, the Shape Game, and the Border Game are presented in the *Kindergarten Psychometric Report* (Najarian et al. 2018). 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.

As can be seen in table 6-3, there was little variation for the *DCCS* pre-switch or post-switch scores at first grade, with mean performance close to a perfect score of 6 out of 6. 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. The Border Game is administered to add difficulty for children who can successfully complete the post-switch items (i.e., those children who achieve a score 5 or 6 on the Shape Game). In the first-grade data collection, some children still had difficulty with the Border Game, as the mean was 4.45 in the fall and 4.62 in the spring out of a possible 6. However, the combined scores for first grade have a mean around 16 out of a possible score of 18, which indicates that some children are reaching ceiling performance on the physical card sort at first grade.

Table 6-3. *Dimensional Change Card Sort* (physical version) variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten and fall and spring first grade: School years 2010–11 and 2011–12

Name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1CSPRES	X1 Card Sort Pre-switch score	15,600	0–6	5.81	0.686
X2CSPRES	X2 Card Sort Pre-switch score	17,150	0–6	5.85	0.632
X3CSPRES	X3 Card Sort Pre-switch score	5,220	0–6	5.91	0.449
X4CSPRES	X4 Card Sort Pre-switch score	15,110	0–6	5.89	0.528
X1CSPSSC	X1 Card Sort Post-switch score	15,600	0–6	5.23	1.679
X2CSPSSC	X2 Card Sort Post-switch score	17,150	0–6	5.55	1.210
X3CSPSSC	X3 Card Sort Post-switch score	5,220	0–6	5.74	0.881
X4CSPSSC	X4 Card Sort Post-switch score	15,110	0–6	5.75	0.883
X1CSBGSC	X1 Card Sort Border Game score	13,280	0–6	3.70	1.185
X2CSBGSC	X2 Card Sort Border Game score	15,690	0–6	4.10	1.314
X3CSBGSC	X3 Card Sort Border Game score	4,930	0–6	4.45	1.326
X4CSBGSC	X4 Card Sort Border Game score	14,430	0–6	4.62	1.311
X1DCCSTOT	X1 Card Sort Combined score	15,600	0–18	14.18	3.343
X2DCCSTOT	X2 Card Sort Combined score	17,150	0–18	15.14	2.815
X3DCCSTOT	X3 Card Sort Combined score	5,220	0–18	15.89	2.293
X4DCCSTOT	X4 Card Sort Combined score	15,110	0–18	16.05	2.347

NOTE: The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (1 for the fall kindergarten round, 2 for the spring kindergarten round, 3 for the fall first-grade round, and 4 for the spring first-grade round). X1 and X2 estimates are weighted by W1C0. X3 estimates are weighted by W3CF3P_30 and X4 estimates are weighted by W4CS4P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted sample *n* indicates the rounded number of cases with valid data regardless of the presence of a valid analytic weight. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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, spring 2011, fall 2011, and spring 2012.

Table 6-4 presents the mean combined/total *DCCS* scores for fall of kindergarten (X1DCCSTOT) and spring of kindergarten (X2DCCSTOT) by selected child characteristics. As stated above, the kindergarten scores are presented here because the combined/total scores were not included on the kindergarten data file and, therefore, they were not included in the *ECLS-K:2011 Kindergarten Psychometric Report* (Najarian et al. 2018). The kindergarten scores for the Color Game, the Shape Game, and the Border Game are presented in the *Kindergarten Psychometric Report* (Najarian et al. 2018).

Table 6-4. Mean *Dimensional Change Card Sort* (physical version) total score for fall and spring kindergarten, by data collection round and child characteristics: School year 2010–11

	Fall 2010 – kindergarten			Spring 2011 – kindergarten		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	15,600	14.18	3.343	17,150	15.14	2.815
Sex						
Male	7,940	14.02	3.466	8,730	14.99	2.942
Female	7,620	14.35	3.193	8,390	15.32	2.654
Race/ethnicity						
White, non-Hispanic	7,590	14.75	2.875	8,080	15.54	2.548
Black, non-Hispanic	2,100	13.38	3.778	2,230	14.36	3.190
Hispanic	3,790	13.40	3.732	4,350	14.69	2.987
Asian, non-Hispanic	1,130	14.23	3.340	1,410	15.21	2.835
Hawaiian, Other Pacific Islander, non-Hispanic	90	13.09	4.274	110	14.08	3.652
American Indian/Alaska Native, non-Hispanic	140	13.63	3.891	150	15.38	2.808
More than one race, non-Hispanic	720	14.59	2.942	780	15.46	2.519
School type						
Public school	13,530	14.14	3.357	15,000	15.09	2.854
Private school	2,080	14.44	3.212	2,130	15.55	2.440

NOTE: Estimates are weighted by W1C0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 0 to 18. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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.

Mean fall first-grade and spring first-grade *DCCS* scores by selected child characteristics are presented in tables 6-5, 6-6, 6-7, and 6-8. Data for the three component scores of the first-grade *DCCS* (pre-switch, post-switch, and Border Game) are presented first, followed by tables with the total scores for first grade.

Table 6-5. Mean *Dimensional Change Card Sort* (physical version) pre-switch score, by data collection round and child characteristics: School year 2011–12

	Fall 2011 – first grade			Spring 2012 – first grade		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	5,220	5.91	0.449	15,110	5.89	0.528
Sex						
Male	2,730	5.89	0.490	7,700	5.88	0.572
Female	2,490	5.92	0.399	7,390	5.91	0.477
Race/ethnicity						
White, non-Hispanic	1,940	5.91	0.452	7,100	5.91	0.511
Black, non-Hispanic	540	5.87	0.540	1,800	5.87	0.597
Hispanic	2,000	5.93	0.402	3,980	5.88	0.547
Asian, non-Hispanic	400	5.90	0.406	1,300	5.92	0.429
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	100	5.99	0.117
American Indian/Alaska Native, non-Hispanic	100	5.88	0.419	130	5.89	0.434
More than one race, non-Hispanic	210	5.94	0.346	680	5.88	0.537
School type						
Public school	4,790	5.90	0.455	13,600	5.89	0.533
Private school	430	5.93	0.386	1,490	5.89	0.480

‡ Reporting standards not met; too few cases for analysis.
NOTE: Fall first-grade estimates are weighted by W3CF3P_30 and spring first-grade estimates are weighted by W4CS4P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 0 to 6. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011 and spring 2012.

Table 6-6. Mean *Dimensional Change Card Sort* (physical version) post-switch score, by data collection round and child characteristics: School year 2011–12

	Fall 2011 – first grade			Spring 2012 – first grade		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	5,220	5.74	0.881	15,110	5.75	0.883
Sex						
Male	2,730	5.71	0.936	7,700	5.74	0.897
Female	2,490	5.78	0.817	7,390	5.76	0.867
Race/ethnicity						
White, non-Hispanic	1,940	5.81	0.698	7,100	5.80	0.724
Black, non-Hispanic	540	5.68	1.040	1,800	5.58	1.236
Hispanic	2,000	5.72	0.955	3,980	5.72	0.934
Asian, non-Hispanic	400	5.67	0.961	1,300	5.74	0.954
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	100	5.81	0.988
American Indian/Alaska Native, non-Hispanic	100	5.34	1.479	130	5.86	0.554
More than one race, non-Hispanic	210	5.61	1.232	680	5.74	0.897
School type						
Public school	4,790	5.74	0.895	13,600	5.75	0.877
Private school	430	5.80	0.721	1,490	5.75	0.875

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

NOTE: Fall first-grade estimates are weighted by W3CF3P_30 and spring first-grade estimates are weighted by W4CS4P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 0 to 6. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011 and spring 2012.

Table 6-7. Mean *Dimensional Change Card Sort* (physical version) Border Game score, by data collection round and child characteristics: School year 2011–12

	Fall 2011 – first grade			Spring 2012 – first grade		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	4,930	4.45	1.326	14,430	4.62	1.311
Sex						
Male	2,560	4.39	1.339	7,320	4.57	1.330
Female	2,370	4.51	1.309	7,080	4.67	1.289
Race/ethnicity						
White, non-Hispanic	1,870	4.70	1.275	6,860	4.83	1.244
Black, non-Hispanic	490	3.94	1.253	1,670	4.20	1.300
Hispanic	1,880	4.14	1.335	3,760	4.37	1.374
Asian, non-Hispanic	380	4.47	1.372	1,250	4.66	1.283
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	100	4.82	1.199
American Indian/Alaska Native, non-Hispanic	90	4.76	1.176	120	4.68	1.304
More than one race, non-Hispanic	200	4.59	1.312	650	4.74	1.254
School type						
Public school	4,510	4.42	1.338	12,960	4.61	1.316
Private school	420	4.73	1.164	1,440	4.77	1.263

‡ Reporting standards not met; too few cases for analysis.
NOTE: Fall first-grade estimates are weighted by W3CF3P_30 and spring first-grade estimates are weighted by W4CS4P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 0 to 6. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011 and spring 2012.

Table 6-8. Mean *Dimensional Change Card Sort* (physical version) total score, by data collection round and child characteristics: School year 2011–12

	Fall 2011 – first grade			Spring 2012 – first grade		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	5,220	15.89	2.293	15,110	16.05	2.347
Sex						
Male	2,730	15.77	2.411	7,700	15.97	2.386
Female	2,490	16.02	2.154	7,390	16.13	2.302
Race/ethnicity						
White, non-Hispanic	1,940	16.25	2.074	7,100	16.37	2.124
Black, non-Hispanic	540	15.27	2.465	1,800	15.32	2.767
Hispanic	2,000	15.56	2.314	3,980	15.72	2.437
Asian, non-Hispanic	400	15.79	2.409	1,300	16.13	2.338
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	100	16.45	2.179
American Indian/Alaska Native, non-Hispanic	100	15.46	3.160	130	16.34	1.794
More than one race, non-Hispanic	210	15.80	2.702	680	16.15	2.349
School type						
Public school	4,790	15.85	2.329	13,600	16.03	2.355
Private school	430	16.36	1.852	1,490	16.22	2.203

‡ Reporting standards not met; too few cases for analysis.
NOTE: Fall first-grade estimates are weighted by W3CF3P_30 and spring first-grade estimates are weighted by W4CS4P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 0 to 18. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011 and spring 2012.

Tables 6-9 and 6-10 present the variable names, descriptions, value ranges, weighted means, and standard deviations for the second-grade *DCCS* total scores. As discussed above, the second-grade *DCCS* total score, which ranges from 0-10, is derived from a different administration of the *DCCS* and is computed differently than the kindergarten- and first-grade *DCCS* scores, which range from 0-18; therefore, the second-grade scores cannot be directly compared to the kindergarten and first-grade scores. For more information on the administration procedures and the scores for both versions of the *DCCS*, see the previous section.

Table 6-9. *Dimensional Change Card Sort* (computerized version) variable names, descriptions, value ranges, weighted means, and standard deviations, second grade: School year 2012–13

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X5DCCSSCR	X5 DCCS Computed (Total) score	4,710	0-10	6.37	1.402
X6DCCSSCR	X6 DCCS Computed (Total) score	13,770	0-10	6.69	1.345

NOTE: The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (5 for the fall second-grade round and 6 for the spring second-grade round). X5 estimates are weighted by W6CF6P_2A0 and X6 estimates are weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted sample *n* indicates the rounded number of cases with valid data regardless of the presence of a valid analytic weight. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012 and spring 2013.

Table 6-10. Mean *Dimensional Change Card Sort* (computerized version) total score, by data collection round and child characteristics: School year 2012–13

	Fall 2012 – second grade			Spring 2013 – second grade		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	4,710	6.37	1.402	13,770	6.69	1.345
Sex						
Male	2,450	6.26	1.492	7,000	6.60	1.424
Female	2,260	6.49	1.290	6,750	6.79	1.249
Race/ethnicity						
White, non-Hispanic	1,750	6.57	1.222	6,510	6.85	1.228
Black, non-Hispanic	460	5.87	1.671	1,570	6.16	1.630
Hispanic	1,840	6.13	1.559	3,720	6.59	1.358
Asian, non-Hispanic	360	6.65	1.200	1,190	6.78	1.325
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	80	6.49	1.528
American Indian/Alaska Native, non-Hispanic	80	6.60	1.097	110	7.04	1.112
More than one race, non-Hispanic	180	6.46	1.274	590	6.88	1.200
School type						
Public school	4,350	6.34	1.413	12,460	6.67	1.355
Private school	340	6.63	1.292	1,270	6.88	1.244

‡ Reporting standards not met; too few cases for analysis.
NOTE: Fall second-grade estimates are weighted by W6CF6P_2A0 and spring second-grade estimates are weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 0 to 10. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012 and spring 2013.

6.1.2 Using the Computerized Version of the *DCCS* for Children Age 8 and Older

As discussed previously, beginning in second grade the 8-year-and-older version of the *DCCS* developed for the NIH Toolbox was administered to all ECLS-K:2011 children regardless of their age on the day of assessment. While the NIH Toolbox *DCCS* assessment is appropriate for ages 3 through 85, children ages 3 through 7 start at a different point than children ages 8 and older. The decision to have two different start points for the task was made by the NIH Toolbox development team primarily to reduce task administration time (Slotkin, Nowinski et al. 2012; Weintraub et al. 2013). The Toolbox *DCCS* consists of pre-switch, post-switch, and mixed-block trials. As described above, the NIH Toolbox administration has children under age 8 begin with the pre-switch trials; children age 8 and older begin with the mixed-block trials as long as they have successfully completed a set of practice trials. This start point transition was selected for the NIH Toolbox *DCCS* based on development work on the NIH Toolbox project that indicated that most 8-year-olds could successfully complete the post-switch trials. However, the age of the start point transition for the NIH Toolbox standard protocol is somewhat arbitrary; a different age start point transition could have been selected and could be appropriate as there is no expectation that a firm developmental change occurs precisely at age 8.

The ECLS-K:2011 administered the 8-year-and-older version of the *DCCS* to all children in the fall and spring data collections beginning with second grade, regardless of the children's age, to simplify task administration and to maximize the comparability of scoring and performance from second grade on. Consequently, all of the children were automatically awarded credit (10 points) for accurately completing all of the pre-switch and post-switch trials that are administered to children ages 3 to 7 in the standard protocol for the NIH Toolbox *DCCS*. Approximately 90 percent of the ECLS-K:2011 children who have a score on the *DCCS* in the fall of second grade (round 5) and approximately 40 percent of the children with a *DCCS* score in the spring of second grade (round 6) were not yet 8 years old when the *DCCS* was administered.

While administering the task in a comparable way using the same start point for all children was done to ensure comparability in administration and scoring, a consequence of the decision to use the 8-year-and-older version for all children beginning in second grade is that more differentiated performance at the lower end of the distribution of scores that could have been captured had the pre-switch and post-switch trials been administered to all children was not. This is true both for children 7 years old and younger as well as for children 8 years old and over who would not have scored with 100 percent accuracy on the pre-switch and post-switch trials.[12] Thus, for some children, regardless of age, their *DCCS* performance in second grade may be overestimated because the score includes points automatically awarded for the pre-switch and post-switch trials as a result of starting with the mixed-block trials. As children get older, this potential overestimation would be expected to diminish as performance improves. Another consideration stemming from the decision to use the 8-year-old and older version for all study children is that those who were 7 years old or younger may have higher average scores than their same-age counterparts in other studies that use the standard NIH Toolbox *DCCS* protocol.

While the NIH Toolbox computerized version of the *DCCS* was not used in the kindergarten and first-grade rounds of the ECLS-K:2011 because it was still under development, it has been administered to children age 3 and older as part of instrument development work. The *NIH Toolbox Technical Manual* (Slotkin, Kallen et al. 2012) presents data on performance on the computerized version of the *DCCS* from age 3 through age 85 that were collected in the Toolbox norming study. Performance of 8-year-old children in the ECLS-K:2011 is similar to performance of 8-year-old children who participated in the NIH Toolbox norming study, as reported in the *Technical Manual*. Eight-year-olds in the norming study of the NIH Toolbox *DCCS* had a mean score of 6.67, compared to an unweighted mean score of 6.41 for ECLS-K:2011 8-year olds in the fall of second grade and an unweighted mean score of 6.81 in the spring of second grade. Performance of ECLS-K:2011 7-year-olds is lower than performance of both the ECLS-K:2011 and NIH Toolbox 8-year-olds, as would be expected based on the points made above. However, performance of the ECLS-K:2011 7-year-olds is higher than the performance of 7-year-olds participating in the NIH Toolbox norming study. The unweighted mean score for ECLS-K:2011 7-year-olds is 6.30 in the fall of second grade and 6.54 in the spring of second grade, compared to a mean of 4.84 for 7-year-olds in the norming study of the NIH Toolbox. It seems likely that ECLS-K:2011 7-year-olds scored higher than NIH Toolbox 7-year-olds at least partially because they automatically received credit for perfect performance on the pre-switch and post-switch scores. Additionally, tables 1 and 2 in the NIH Toolbox *Technical Manual* present data that show a larger increase in scores from age 7 to age 8 than is seen between other ages. It is possible that the larger magnitude of the difference between these ages is the result of the change in start point on the *DCCS* that occurs at age 8, when all children begin to automatically receive the maximum number of pre-switch and post-switch accuracy points. In other

[12] In the version for 3 to 7 year olds, a child will end the *DCCS* task after the post-switch trials if he or she has not scored enough items correctly. A child must correctly answer at least 4 of the 5 pre-switch items and 4 of the 5 post-switch items to be administered the mixed-block trials. In instances where a child answered only 4 rather than all 5 of either or both the pre-switch or post-switch items correctly, scores produced from the 3 to 7 year version would differ from those produced from the 8-year-and-older version, even if scores on the mixed-block trials were identical. The result of this scoring is that more differentiated performance at the lower end of the distribution would be captured by the 3 to 7-year-old version than by the 8-year-and-older version of the task.

words, the higher scores at age 8 may be the result of an artifact in the scoring rather than a true increase in performance on the *DCCS* from age 7 to age 8. However, data from the validation study of the NIH Toolbox and from the norming study of the NIH Toolbox are not sufficient to determine the extent to which this pattern is an artifact of scoring and the extent to which this pattern is a result of developmental change.

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 up to 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 presented with up to 5 three-number sequences. The sequence becomes increasingly longer, up to a maximum sequence 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 kindergarten–second grade data file. The maximum number of items any child could have been administered in the fall or spring first-grade or second-grade 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 first-grade assessments begin with "C3," spring first-grade assessments begin with "C4," fall second-grade assessments begin with "C5," and spring second-grade assessments begin with "C6." Variable descriptions for these items indicate the length of the digit sequence (e.g., C4 Numbers Reversed Two-Digit Sequence #1).

As in kindergarten, in first grade the Numbers Reversed subtask was administered in Spanish for children routed through the assessment in Spanish. Data from English and Spanish administrations are combined into the same variables. Researchers who want to account for language of administration in their analyses can use the variables X3FLSCRN and X4FLSCRN, which are also in the data file, to identify which cases were administered Numbers Reversed in English and which cases were administered it in Spanish in first grade. Numbers Reversed was administered only in English in second grade; therefore, there is no flag variable indicating language of administration for second grade.

In addition to the item-level data, five scores developed using guidelines from the publisher scoring materials are included in the data file for Numbers Reversed: the *W*-ability score, the age standard score, the grade standard score, the age percentile score, and the grade percentile score. The grade standard score and the grade percentile score are additional scores for all rounds of data collection being released for the first time on the kindergarten-second grade data file.

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. 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 age or grade percentile rank and/or age or grade 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.[13]

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 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

[13] 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).

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.

Researchers should keep in mind that in first grade (as in kindergarten) the Numbers Reversed subtask was administered in both English and Spanish. 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. As mentioned above, researchers who want to account for language of administration in their analyses can use the data flag provided on the data file for each round (X*FLSCRN) to identify which cases were administered

Numbers Reversed in English and which cases were administered Numbers Reversed in Spanish. All children were administered the assessments in English in the second-grade collections. Therefore, the second-grade Numbers Reversed scores for all children are based on an English administration of the assessment, and data flags to indicate language administration in second grade (round 5 and 6) are not provided on the data file.[14]

Although the *W* score is reflective of the average performance of 10-year-olds, and the ECLS-K:2011 children were in first and second grade during collection, it is included in the data file 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 6 or 7 years old during the first-grade data collections and 7 or 8 years old during the second-grade data collections while 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. As expected, the discrepancy is declining as the participating children grow older.

A score of 403 (393 for the Spanish administration) 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 raw score of 0 (which translates to a *W* score of 403 for the English administration and 393 for the Spanish administration) 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 the fall of first grade, less than 13 percent of the children scored at the lowest scalable score, and less than 6 percent scored at the lowest scalable score in the spring of first grade. In the fall of second grade, less than 4 percent scored the lowest scalable score, and slightly more than 2 percent received the lowest score in the spring. These percentages show a general improvement over time.

[14] More information about how children's home language affected children's routing through the assessment battery in each round of data collection is provided in chapter 5.

Another factor that may contribute to the number of children scoring 403 (and 393 for Spanish), especially in kindergarten, 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 fully determine the extent to which improper administration of the practice items affected the results, and improvements to training and task administration were implemented between kindergarten and first grade.[15] However, researchers should keep in mind that this may have affected performance for some (but not all) children. In analyzing the data, researchers need to decide how to handle the 403 (393 for Spanish) scores; the decision for how to do so is left up to the analyst based on his or her analytic goals.

The four additional Numbers Reversed scores are the *age standard score,* the *grade standard score,* the *age percentile score,* and the *grade percentile score.* These scores indicate children's status relative to their peers through age-normed and grade-normed transformations of the data. That is, these scores are relative to *same-aged* or *same-grade* subjects in the WJ III norming sample. The standard scores are created by the publisher and have 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 scores describe 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 or grade as the ECLS-K:2011 subjects.

As with the first-grade *W* score, the publisher's scoring protocols result in age standard and age percentile scores for first grade 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

[15] During the kindergarten data collections, all the practice items appeared on a single screen in the computer program used to administer Numbers Reversed, and the assessor had to determine when to administer additional practice items and when to begin the test trials. For the first grade data collections, the computer screens were changed so that each practice item appeared on a separate screen and the computer determined which item to administer next based on the child's answer (correct or incorrect) for each practice item. These changes better ensured that the practice items were administered in a standardized manner and that children received the appropriate number of practice items before moving on to the test trials.

score, regardless of their raw score. Again, there are variables in the data file indicating language of administration in first grade (X3FLSCRN and X4FLSCRN) that analysts may want to include in their analytic models.

For both the age-normed scores and the grade-normed scores, 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 and spring kindergarten, fall and spring first-grade, and second-grade Numbers Reversed scores are shown in table 6-11. The weighted means for the ECLS-K:2011 population are lower than the established means from the WJ III norming sample:[16] the average *W* scores for the ECLS-K:2011 population are less than 500, the average age and grade standard scores are less than 100, and the average age and grade percentile scores are less than 50.The lower mean for the *W* scores 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. The lower means for the standard and percentile scores in the ECLS-K:2011 may also be attributable to differences between the ECLS-K:2011 population and the WJ III norming sample.

[16] Normative data for the WJ III were gathered from 8,818 subjects in more than 100 geographically diverse U.S. communities (McGrew and Woodcock 2001). 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; and occupation of adults in the labor force.

Table 6-11. Numbers Reversed variable names, descriptions, value ranges, weighted means, and standard deviations: School years 2010–11, 2011–12 and 2012–13

Variable name	Description	*n*	Range of Possible values	Weighted mean	Standard deviation
X1NRWABL	X1 Numbers Reversed *W*-Ability Score	15,600	393-603	432.56	30.028
X1NRSSCR	X1 Numbers Reversed Age Standard Score	14,450	45-200	93.10	16.510
X1NRSSGR	X1 Numbers Reversed Grade Standard Score	15,600	74-190	96.40	14.569
X1NRPERC	X1 Numbers Reversed Age Percentile	14,450	0-100	37.89	31.786
X1NRPEGR	X1 Numbers Reversed Grade Percentile	15,600	0-100	41.98	30.886
X2NRWABL	X2 Numbers Reversed *W*-Ability Score	17,150	393-603	449.49	30.412
X2NRSSCR	X2 Numbers Reversed Age Standard Score	17,120	39-200	94.92	17.017
X2NRSSGR	X2 Numbers Reversed Grade Standard Score	17,150	54-200	94.76	16.049
X2NRPERC	X2 Numbers Reversed Age Percentile	17,120	0-100	42.44	30.970
X2NRPEGR	X2 Numbers Reversed Grade Percentile	17,150	0-100	41.89	29.980
X3NRWABL	X3 Numbers Reversed *W*-Ability Score	5,220	393-603	458.42	27.990
X3NRSSCR	X3 Numbers Reversed Age Standard Score	5,220	36-200	94.21	16.969
X3NRSSGR	X3 Numbers Reversed Grade Standard Score	5,220	29-200	95.19	17.815
X3NRPERC	X3 Numbers Reversed Age Percentile	5,220	0-100	41.23	28.832
X3NRPEGR	X3 Numbers Reversed Grade Percentile	5,220	0-100	43.61	29.857
X4NRWABL	X4 Numbers Reversed *W*-Ability Score	15,110	393-603	469.56	25.395
X4NRSSCR	X4 Numbers Reversed Age Standard Score	15,100	32-200	95.90	16.872
X4NRSSGR	X4 Numbers Reversed Grade Standard Score	15,110	19-200	95.42	18.159
X4NRPERC	X4 Numbers Reversed Age Percentile	15,100	0-100	44.35	28.470
X4NRPEGR	X4 Numbers Reversed Grade Percentile	15,110	0-100	44.07	29.276
X5NRWABL	X5 Numbers Reversed *W*-Ability Score	4,730	403-603	473.93	23.736
X5NRSSCR	X5 Numbers Reversed Age Standard Score	4,730	29-200	94.93	16.574
X5NRSSGR	X5 Numbers Reversed Grade Standard Score	4,730	23-200	95.85	17.561
X5NRPERC	X5 Numbers Reversed Age Percentile	4,730	0-100	42.13	27.609
X5NRPEGR	X5 Numbers Reversed Grade Percentile	4,730	0-100	44.17	28.742
X6NRWABL	X6 Numbers Reversed *W*-Ability Score	13,830	403-603	480.70	22.841
X6NRSSCR	X6 Numbers Reversed Age Standard Score	13,830	25-200	95.80	16.749
X6NRSSGR	X6 Numbers Reversed Grade Standard Score	13,830	19-200	95.52	17.715
X6NRPERC	X6 Numbers Reversed Age Percentile	13,830	0-100	43.67	27.765
X6NRPEGR	X6 Numbers Reversed Grade Percentile	13,830	0-100	43.59	28.680

NOTE: The name and description for each variable in the table begin with an “X,” indicating that it is a derived/calculated variable, and a data collection round number (1 for the fall kindergarten round, 2 for the spring kindergarten round, 3 for the fall first-grade round, 4 for the spring first-grade round, 5 for the fall second-grade round, and 6 for the spring second-grade round). X1 and X2 estimates are weighted by W1C0, X3 estimates are weighted by W3CF3P_30, X4 estimates are weighted by W4CS4P_20, X5 estimates are weighted by W6CF6P_2A0, and X6 estimates are weighted by W6CS6P_20. See the ECLS-K:2011 data file User’s Manuals for explanation of the weighting variables. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

Means and standard deviations for the Numbers Reversed scores are provided by data collection round and by child characteristics in tables 6-12 through 6-18. The *W*-ability score, age standard score, grade standard score, age percentile score, and grade percentile scores are provided by child characteristics for the first-grade and second-grade data collections. Because grade standard scores and the grade percentile scores for all rounds of data collection are being released for the first time on the kindergarten-second grade data file, the grade standard score and the grade percentile score for fall and spring of kindergarten are also provided by child characteristics in this report. The fall and spring kindergarten W-ability scores, age standard scores, and age percentile scores are provided by child characteristics in the kindergarten psychometric report, *ECLS-K:2011 Kindergarten Psychometric Report* (Najarian et al. 2018).

Table 6-12. Mean Numbers Reversed grade standard score for fall and spring kindergarten, by child characteristics: School year 2010–11

	Fall 2010 – kindergarten			Spring 2011 – kindergarten		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	15,600	96.40	14.569	17,150	94.76	16.049
Sex						
Male	7,940	95.84	14.547	8,730	94.06	16.519
Female	7,620	97.04	14.576	8,390	95.52	15.499
Race/ethnicity						
White, non-Hispanic	7,590	99.78	14.475	8,080	98.17	14.981
Black, non-Hispanic	2,100	91.76	13.078	2,230	89.58	16.048
Hispanic	3,790	91.00	12.927	4,350	89.52	16.078
Asian, non-Hispanic	1,130	99.62	15.583	1,410	98.52	16.404
Hawaiian, Other Pacific Islander, non-Hispanic	90	95.95	15.799	110	95.47	15.667
American Indian/Alaska Native, non-Hispanic	140	95.24	14.046	150	93.02	16.112
More than one race, non-Hispanic	720	99.32	14.792	780	97.61	15.185
School type						
Public school	13,520	95.76	14.428	15,000	94.14	16.117
Private school	2,080	101.60	14.654	2,130	99.76	14.553

NOTE: Estimates are weighted by W1C0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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.

Table 6-13. Mean Numbers Reversed grade percentile rank for fall and spring kindergarten, by child characteristics: School year 2010–11

	Fall 2010 – kindergarten			Spring 2011 – kindergarten		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	15,600	41.98	30.886	17,150	41.89	29.980
Sex						
Male	7,940	40.71	30.821	8,730	40.79	30.657
Female	7,620	43.40	30.910	8,390	43.10	29.207
Race/ethnicity						
White, non-Hispanic	7,590	49.09	30.671	8,080	48.08	28.902
Black, non-Hispanic	2,100	32.12	28.175	2,230	32.57	28.976
Hispanic	3,790	30.70	27.494	4,350	32.32	28.854
Asian, non-Hispanic	1,130	48.51	32.207	1,410	48.77	30.657
Hawaiian, Other Pacific Islander, non-Hispanic	90	40.09	31.619	110	42.57	29.438
American Indian/Alaska Native, non-Hispanic	140	39.74	30.516	150	39.38	29.749
More than one race, non-Hispanic	720	48.12	31.092	780	46.99	29.277
School type						
Public school	13,520	40.66	30.631	15,000	40.76	29.973
Private school	2,080	52.64	30.867	2,130	51.09	28.414

NOTE: Estimates are weighted by W1C0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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.

Table 6-14. Mean Numbers Reversed *W*-ability score, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall 2011– first grade			Spring 2012 – first grade			Fall 2012 – second grade			Spring 2013 – second grade		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	5,220	458.42	27.990	15,110	469.56	25.395	4,730	473.93	23.736	13,830	480.70	22.841
Sex												
Male	2,730	456.22	29.488	7,700	468.41	26.472	2,450	472.99	24.927	7,040	480.07	23.389
Female	2,490	460.75	26.101	7,390	470.79	24.140	2,270	474.93	22.371	6,780	481.36	22.225
Race/ethnicity												
White, non-Hispanic	1,940	462.81	25.176	7,100	473.74	22.702	1,760	477.12	21.618	6,540	483.47	21.348
Black, non-Hispanic	540	452.55	30.569	1,800	462.08	27.763	460	468.12	26.366	1,570	473.48	25.978
Hispanic	2,000	451.18	29.515	3,980	463.24	27.522	1,860	469.48	24.519	3,730	477.55	23.141
Asian, non-Hispanic	400	465.94	30.375	1,300	476.06	24.650	360	479.58	25.226	1,190	485.88	21.759
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	100	470.89	24.312	‡	‡	‡	80	482.29	19.708
American Indian/Alaska Native, non-Hispanic	100	456.64	26.505	130	470.84	22.440	80	469.49	21.946	110	478.15	20.592
More than one race, non-Hispanic	210	459.96	28.301	680	472.75	23.798	180	474.80	26.160	590	482.77	22.639
School type												
Public school	4,790	457.37	28.026	13,600	468.95	25.555	4,370	473.39	23.839	12,520	480.32	23.040
Private school	430	469.59	25.058	1,490	475.38	23.212	340	480.36	21.661	1,280	485.08	19.801

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

NOTE: Fall first-grade estimates are weighted by W3CF3P_30, spring first-grade estimates are weighted by W4CS4P_20, fall second-grade estimates are weighted by W6CF6P_2A0, and spring second-grade estimates are weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Detail may not sum to total due to rounding and/or missing data. *SD* = standard deviation. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011, spring 2012, fall 2012, and spring 2013.

Table 6-15. Mean Numbers Reversed age standard score, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall 2011–first grade			Spring 2012 – first grade			Fall 2012 – second grade			Spring 2013 – second grade		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	5,220	94.21	16.969	15,100	95.90	16.872	4,730	94.93	16.574	13,830	95.80	16.749
Sex												
Male	2,730	92.72	17.831	7,690	94.95	17.544	2,450	94.10	17.397	7,040	95.19	17.136
Female	2,490	95.79	15.844	7,390	96.91	16.067	2,270	95.79	15.614	6,770	96.45	16.302
Race/ethnicity												
White, non-Hispanic	1,940	96.55	15.463	7,100	98.25	15.594	1,760	96.94	15.433	6,540	97.54	15.931
Black, non-Hispanic	540	90.89	18.371	1,800	91.14	17.782	460	91.02	17.746	1,570	90.78	18.623
Hispanic	2,000	90.18	17.617	3,980	92.31	17.727	1,860	92.03	16.684	3,730	93.82	16.680
Asian, non-Hispanic	400	100.05	19.129	1,300	101.49	17.223	360	99.71	18.506	1,190	100.56	16.303
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	100	96.55	16.609	‡	‡	‡	80	96.47	14.721
American Indian/Alaska Native, non-Hispanic	100	92.26	15.402	130	95.99	15.231	80	91.49	15.356	110	93.57	15.083
More than one race, non-Hispanic	210	95.19	17.672	680	97.97	16.460	180	95.54	18.647	590	97.35	16.869
School type												
Public school	4,790	93.53	16.909	13,590	95.50	16.924	4,370	94.51	16.660	12,510	95.54	16.878
Private school	430	101.47	15.829	1,480	99.69	16.008	340	99.72	14.833	1,280	98.84	14.763

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

NOTE: Fall first-grade estimates are weighted by W3CF3P_30, spring first-grade estimates are weighted by W4CS4P_20, fall second-grade estimates are weighted by W6CF6P_2A0, and spring second-grade estimates are weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. *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), fall 2011, spring 2012, fall 2012, and spring 2013.

Table 6-16. Mean Numbers Reversed grade standard score, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall 2011– first grade			Spring 2012 – first grade			Fall 2012– second grade			Spring 2013 – second grade		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	5,220	95.19	17.815	15,110	95.42	18.159	4,730	95.85	17.561	13,830	95.52	17.715
Sex												
Male	2,730	94.04	18.693	7,700	94.71	18.803	2,450	95.36	18.365	7,040	95.14	18.059
Female	2,490	96.42	16.744	7,390	96.17	17.418	2,270	96.37	16.656	6,780	95.93	17.333
Race/ethnicity												
White, non-Hispanic	1,940	98.15	15.885	7,100	98.49	16.316	1,760	98.27	16.233	6,540	97.64	16.756
Black, non-Hispanic	540	91.20	19.438	1,800	89.97	19.527	460	91.28	19.359	1,570	89.96	19.758
Hispanic	2,000	90.19	18.718	3,980	90.73	19.480	1,860	92.39	17.876	3,730	93.10	17.704
Asian, non-Hispanic	400	100.31	19.932	1,300	100.17	18.446	360	99.92	18.493	1,190	99.63	17.380
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	100	95.93	18.218	‡	‡	‡	80	96.41	15.665
American Indian/Alaska Native, non-Hispanic	100	94.27	16.914	130	96.12	15.422	80	94.32	15.353	110	93.58	15.774
More than one race, non-Hispanic	210	97.20	18.048	680	98.04	17.131	180	97.12	18.822	590	97.26	17.627
School type												
Public school	4,790	94.49	17.837	13,600	94.98	18.230	4,370	95.45	17.636	12,520	95.24	17.837
Private school	430	102.77	15.726	1,490	99.59	17.060	340	100.60	16.025	1,280	98.84	15.736

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

NOTE: Fall first-grade estimates are weighted by W3CF3P_30, spring first-grade estimates are weighted by W4CS4P_20, fall second-grade estimates are weighted by W6CF6P_2A0, and spring second-grade estimates are weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. *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), fall 2011, spring 2012, fall 2012, and spring 2013.

Table 6-17. Mean Numbers Reversed age percentile rank, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall 2011 – first grade			Spring 2012 – first grade			Fall 2012 – second grade			Spring 2013 – second grade		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	5,220	41.23	28.832	15,100	44.35	28.470	4,730	42.13	27.609	13,830	43.67	27.765
Sex												
Male	2,730	39.11	29.560	7,960	42.99	28.801	2,450	41.09	28.153	7,040	42.83	27.682
Female	2,490	43.49	27.857	7,390	45.79	28.043	2,270	43.23	26.981	6,770	44.56	27.824
Race/ethnicity												
White, non-Hispanic	1,940	44.93	27.997	7,100	48.24	27.546	1,760	44.90	27.054	6,540	46.38	27.446
Black, non-Hispanic	540	36.38	28.500	1,800	36.83	28.224	460	36.75	27.086	1,570	36.07	27.430
Hispanic	2,000	34.65	28.571	3,980	38.39	28.471	1,860	37.46	27.104	3,730	40.36	27.138
Asian, non-Hispanic	400	51.14	31.181	1,300	53.58	28.926	360	51.69	30.167	1,190	52.06	28.025
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	100	44.02	29.280	‡	‡	‡	80	44.10	27.518
American Indian/Alaska Native, non-Hispanic	100	37.34	26.177	130	43.48	26.748	80	35.66	26.262	110	38.67	25.779
More than one race, non-Hispanic	210	42.80	29.944	680	47.36	28.442	180	44.80	28.772	590	46.12	29.009
School type												
Public school	4,790	40.08	28.612	13,590	43.68	28.447	4,370	41.44	27.512	12,510	43.27	27.741
Private school	430	53.66	28.026	1,480	50.79	27.947	340	50.27	27.308	1,280	48.30	27.314

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

NOTE: Fall first-grade estimates are weighted by W3CF3P_30, spring first-grade estimates are weighted by W4CS4P_20, fall second-grade estimates are weighted by W6CF6P_2A0, and spring second-grade estimates are weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. *S D* = 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), fall 2011, spring 2012, fall 2012, and spring 2013.

Table 6-18. Mean Numbers Reversed grade percentile rank, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall 2011 – first grade			Spring 2012 – first grade			Fall 2012 – second grade			Spring 2013 – second grade		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	5,220	43.61	29.857	15,110	44.07	29.276	4,730	44.17	28.742	13,830	43.59	28.680
Sex												
Male	2,730	42.06	30.703	7,700	43.22	29.586	2,450	43.71	29.297	7,040	43.18	28.560
Female	2,490	45.25	28.838	7,390	44.98	28.916	2,270	44.65	28.137	6,780	44.03	28.800
Race/ethnicity												
White, non-Hispanic	1,940	48.11	28.733	7,100	48.74	28.263	1,760	47.36	28.182	6,540	46.86	28.416
Black, non-Hispanic	540	38.10	29.754	1,800	36.21	29.038	460	38.32	28.315	1,570	35.41	28.119
Hispanic	2,000	35.46	29.621	3,980	36.90	29.072	1,860	38.79	28.415	3,730	39.55	27.907
Asian, non-Hispanic	400	52.40	30.981	1,300	51.32	29.773	360	52.14	30.256	1,190	50.73	29.032
Hawaiian, Other Pacific Islander, non-Hispanic	‡	‡	‡	100	44.19	30.702	‡	‡	‡	80	44.69	28.391
American Indian/Alaska Native, non-Hispanic	100	42.12	28.332	130	43.29	26.641	80	40.96	27.722	110	38.90	26.584
More than one race, non-Hispanic	210	46.77	30.922	680	47.69	29.087	180	47.68	29.064	590	46.43	29.554
School type												
Public school	4,790	42.44	29.733	13,600	43.37	29.285	4,370	43.48	28.694	12,520	43.16	28.633
Private school	430	56.18	28.084	1,490	50.87	28.404	340	52.36	27.961	1,280	48.61	28.508

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

NOTE: Fall first-grade estimates are weighted by W3CF3P_30, spring first-grade estimates are weighted by W4CS4P_20, fall second-grade estimates are weighted by W6CF6P_2A0, and spring second-grade estimates are weighted by W6CS6P_20. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Detail may not sum to total due to rounding and/or missing data. *SD* = standard deviation. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011, spring 2012, fall 2012, and spring 2013.

7. PSYCHOMETRIC CHARACTERISTICS OF THE INDIRECT MEASURES

In the ECLS-K:2011, parent and teacher reports about children's skills and behaviors are referred to as indirect measures. This chapter describes the indirect measures included in the first- and second-grade rounds of data collection. For information about the indirect measures used in kindergarten, see chapter 3 of the *User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic Codebook, Public Version* (NCES 2015-074) (Tourangeau et al. 2015). In first grade, parents and teachers of participating students completed indirect measures of the students' social skills. In first grade, teachers also completed indirect measures of students' executive functioning skills and a measure of the teacher's perceived relationship with the student. In second grade, teachers completed indirect measures of students' social skills and executive functioning and the student-teacher relationship. There were no parent-reported indirect measures of children's social skills or behavior included in the second-grade collections.

Two versions of the teacher-level and child-level teacher questionnaires were used in the spring of first grade: one version for students who were in first grade or higher during the data collection period and one for students who had been retained in kindergarten for the 2011–12 school year. Details of the differences in these questionnaires are presented in chapter 2 of the *User's Manual for the ECLS-K:2011 Kindergarten–First Grade Data File and Electronic Codebook, Public Version* (NCES 2015-078) (Tourangeau et al. 2015). In the tables below, the variables applicable to first-grade students[1] begin with X4, while those for children retained in kindergarten begin with X4K. For second grade, different versions of the questionnaires were not used; all teachers were asked to complete the same survey regardless of a student's grade level. In the tables below, the variables applicable to the fall second-grade and spring second-grade rounds of collection begin with X5 and X6, respectively.

[1] For ease of presentation, this chapter refers to all students who were not retained in kindergarten in the 2011–12 school year as "first-grade students," although the reader should keep in mind that a very small number of students had been advanced to a higher grade and are included in the estimates for the "first-grade students."

7.1 Teacher Measures

7.1.1 *Children's Behavior Questionnaire* (*CBQ*) and *Temperament in Middle Childhood Questionnaire* (*TMCQ*)

The spring first-grade child-level teacher questionnaires (both the version for students in first grade and the version for students in kindergarten) included 12 items from the *Short Form of the Children's Behavior Questionnaire* (*CBQ*; Putnam and Rothbart 2006)[2] 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 related to executive functioning. Rothbart describes inhibitory control as the "capacity to plan and to suppress inappropriate approach responses under instructions or in novel or uncertain situations" (Rothbart et al. 2001, p. 1406). Attentional focusing is described as the "capacity to maintain attentional focus on task-related channels. 'When picking up toys or other jobs, usually keeps at the task until it's done.'" (Rothbart et al. 2001, p. 1406). 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 *CBQ* is appropriate for assessment of children ages 3 through 7 years, so it could not be used past the first-grade rounds of data collection. To remain age appropriate, the *CBQ* was replaced with the *Temperament in Middle Childhood Questionnaire*[3] (*TMCQ*; Simonds and Rothbart 2004) in the spring of second grade. The *TMCQ* was designed as an upward age-extension of the *CBQ* and is appropriate for ages 7 through 10 years. While many of the items from the *TMCQ* are different from the items on the *CBQ*, the items are believed to assess the same or similar constructs in an age-appropriate way. Teachers received the same instructions for the *CBQ* and *TMCQ* items; however, the *TMCQ* items were rated on a 5-point scale instead of the 7-point scale used for the *CBQ* items. For the *TMCQ* items, teachers used a 5-point scale ranging from "almost always untrue" to "almost always true," with a middle option of "sometimes true, sometimes untrue." As in the *CBQ*, there was a "not applicable" option that the teacher could select if the statement or situation did not apply to the child.

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

[3] The *Temperament in Middle Childhood Questionnaire* is a copyrighted instrument and has been used with permission. Adapted from the *Temperament in Middle Childhood Questionnaire.* © 2004 Jennifer Simonds and Mary K. Rothbart, University of Oregon. Used with permission.

Kindergarten, first-grade, and second-grade item-level data for the items that make up the attentional focusing and inhibitory control scales are provided on the kindergarten-second grade data file. Item-level data for these *CBQ* and *TMCQ* scales are provided for the first time on the kindergarten-second grade data file because the study recently received copyright permission to include these data on ECLS-K:2011 data files. Variable names for the item-level data for the fall and spring kindergarten child-level teacher questionnaires begin with “T1” and “T2,” respectively. Variable names for the item-level data from the spring first-grade child-level teacher questionnaire for children in first grade begin with “T4,” while variable names for children held back in kindergarten begin with “T4K.” Variable names for the item-level data from the spring second-grade data collection begin with “T6.”

The data file includes two scale scores for each round of data collection in which this measure was included: (1) attentional focus and (2) inhibitory control. In first grade these scores are derived from the *CBQ* and in second grade these scores are derived primarily from the *TMCQ*, as explained further below. The scale scores were developed based on guidelines from the instrument developers for both the *CBQ* and *TMCQ*.

In first grade, the ECLS-K:2011 fielded all six items from the Attentional Focusing subscale and all six items from the Inhibitory Control subscale of the *CBQ Short Form*. As such, the first-grade attentional focus and inhibitory control scores are each based on all six items in the relevant *Short Form* subscale. Because the *CBQ* was initially designed as a parent-report measure, the item wording for three of the items from the *CBQ* Inhibitory Control subscale was modified for use in the ECLS-K:2011 to make them more appropriate for a school setting.

In second grade, the ECLS-K:2011 fielded six of the seven items from the *TMCQ* Attentional Focusing subscale. For the inhibitory control dimension, the ECLS-K:2011 fielded six of the eight items from the *TMCQ* Inhibitory Control subscale and one item from the *CBQ* Inhibitory Control subscale. Therefore, the second-grade attentional focusing scale score reflects the six items fielded by the ECLS-K:2011, not the full set of items in the original *TMCQ* scale. The second-grade inhibitory control scale score reflects the seven items fielded by the ECLS-K:2011 (six from the *TMCQ* and one from the *CBQ*), again not the full set of items in the original *TMCQ* scale. Because the *TMCQ* was designed as a parent-report measure, the item wording on one item from the Attentional Focusing subscale was modified slightly to make it more appropriate for a school setting and one item on the Inhibitory Control subscale was similarly modified.

For both the first- and second-grade attention focusing and inhibitory control scales, 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 or seven items that made up the scale. Scores on rated items were summed and divided by the number of items rated to derive the scale score. 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 a task. Higher scale scores on the inhibitory control scale indicate that the child exhibited more behaviors that demonstrate the ability to hold back or suppress a behavior as appropriate for instructions given or in novel or uncertain situations (Rothbart et al. 2001). The variable names, descriptions, value ranges, weighted means, and standard deviations for these scales are shown in tables 7-1 and 7-2.

Table 7-3 presents the internal consistency reliability coefficients (Cronbach's alpha) for the teacher-reported attentional focus and inhibitory control scales for first and second grades. For round 4 (the spring of first grade), the attentional focus scale has an internal consistency reliability coefficient (Cronbach's alpha) of .83 for children in first grade (X4ATTNFS) and .86 for children retained in kindergarten (X4KATTNFS), and the inhibitory control scale has an internal consistency reliability coefficient (Cronbach's alpha) of .86 for both children in first grade (X4INBCNT) and those retained in kindergarten (X4KINBCNT). For round 6 (the spring of second grade), the attentional focus scale (X6ATTMCQ) has an internal consistency reliability coefficient (Cronbach's alpha) of .96, and the inhibitory control scale (X6INTMCQ) has an internal consistency reliability coefficient (Cronbach's alpha) of .87.

Table 7-1. *Children's Behavior Questionnaire* variable names, descriptions, value ranges, weighted means, and standard deviations: Spring 2012

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X4ATTNFS	X4 Teacher Report Attentional Focus	13,390	1–7	4.84	1.292
X4INBCNT	X4 Teacher Report Inhibitory Control	13,400	1–7	5.04	1.287
X4KATTNFS	X4K Teacher Report Attentional Focus	420	1–7	4.61	1.323
X4KINBCNT	X4K Teacher Report Inhibitory Control	420	1–7	4.88	1.267

NOTE: Items contributing to these scales come from the *Children's Behavior Questionnaire* (Putnam and Rothbart 2006). Estimates weighted by W4CS4P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

Table 7-2. *Temperament in Middle Childhood Questionnaire* variable names, descriptions, value ranges, weighted means, and standard deviations: Spring 2013

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X6ATTMCQ	X6 TMCQ Teacher Report Attentional Focus	12,660	1–5	3.47	1.122
X6INTMCQ	X6 TMCQ Teacher Report Inhibitory Control	12,660	1–5	3.67	0.845

NOTE: Items contributing to these scales come from the *Temperament in Middle Childhood Questionnaire* (Simonds and Rothbart 2004). Estimates weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2013.

Table 7-3. Reliability estimates for the teacher-reported attentional focus and inhibitory control: Spring 2012 and spring 2013

Variable name	Description	Number of items	Reliability coefficient
X4ATTNFS	X4 Teacher Report Attentional Focus	6	.83
X4INBCNT	X4 Teacher Report Inhibitory Control	6	.86
X4KATTNFS	X4 Teacher Report Attentional Focus	6	.86
X4KINBCNT	X4 Teacher Report Inhibitory Control	6	.86
X6ATTMCQ	X6 TMCQ Teacher Report Attentional Focus	6	.96
X6INTMCQ	X6 TMCQ Teacher Report Inhibitory Control	7	.87

NOTE: Items contributing to these scales come from the *Children's Behavior Questionnaire* (Putnam and Rothbart 2006) and the *Temperament in Middle Childhood Questionnaire* (Simonds and Rothbart 2004). The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (4 for the spring first-grade round and 6 for the spring second-grade round). X4 estimates weighted by W4CS4P_2T0. X6 estimates weighted by W6CS6P_2T0.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2012 and spring 2013.

Correlations between the attentional focusing and inhibitory control scores produced from the *CBQ* items in first grade (X4ATTNFS and X4INBCNT) and those scores produced from the *TMCQ* items in second grade (X6ATTMCQ and X6INTMCQ) are presented in table 7-4. The correlation between attentional focusing measured in the spring of first grade using items from the *CBQ* and in the spring of second grade using items from the *TMCQ* was .56 (p < .001). The correlation between inhibitory control measured in the spring of first grade using items from the *CBQ* and in the spring of second grade using items from the *TMCQ* was .56 (p < .001). While these are moderate correlations, it is important to remember that different sets of items were used to derive these scores. For this reason, kindergarten and first-grade scores should not be directly compared to scores from second grade for the purpose of measuring change over time.

Table 7-4. Intercorrelations among attentional focusing and inhibitory control scales in first and second grades: Spring 2012 and spring 2013

	Spring first grade		Spring second grade	
Measures	Attentional Focus (X4ATTNFS)	Inhibitory Control (X4INBCNT)	Attentional Focus (X6ATTMCQ)	Inhibitory Control (X6INTMCQ)
X4ATTNFS	1.00			
X4INBCNT	.77	1.00		
X6ATTMCQ	.56	.57	1.00	
X6INTMCQ	.49	.56	.77	1.00

NOTE: Items contributing to these scales come from the *Children's Behavior Questionnaire* (Putnam and Rothbart 2006) and the *Temperament in Middle Childhood Questionnaire* (Simonds and Rothbart 2004). The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (4 for the spring first-grade round and 6 for the spring second-grade round). X4 estimates weighted by W4CS4P_2T0. X6 estimates weighted by W6CS6P_2T0. All correlations $p < .001$. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012 and spring 2013.

Mean scores for the teacher-reported attentional focus and inhibitory control scales are presented by child characteristics in tables 7-5 (spring of first grade) and 7-6 (spring of second grade).

Table 7-5. Mean *Children's Behavior Questionnaire* attentional focus and inhibitory control scores, by child characteristics, spring first grade: Spring 2012

	Spring first grade					
	Attentional Focus (X4ATTNFS)			Inhibitory Control (X4INBCNT)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	13,390	4.84	1.292	13,400	5.04	1.287
Sex						
Male	6,770	4.57	1.305	6,770	4.72	1.316
Female	6,600	5.12	1.215	6,610	5.37	1.165
Race/ethnicity						
White, non-Hispanic	6,480	4.90	1.293	6,480	5.10	1.290
Black, non-Hispanic	1,600	4.51	1.311	1,600	4.68	1.330
Hispanic	3,420	4.83	1.266	3,420	5.05	1.256
Asian, non-Hispanic	1,090	5.32	1.175	1,090	5.44	1.146
Hawaiian, Other Pacific Islander, non-Hispanic	90	4.91	1.195	90	5.01	1.282
American Indian/Alaska Native, non-Hispanic	120	4.53	1.234	120	5.11	1.137
More than one race, non-Hispanic	590	4.87	1.258	590	5.05	1.222
School type						
Public school	11,970	4.83	1.296	11,990	5.03	1.288
Private school	1,410	4.96	1.239	1,400	5.16	1.263

NOTE: Items contributing to these scales come from the *Children's Behavior Questionnaire* (Putnam and Rothbart 2006). Estimates weighted by W4CS4P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 7. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. *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), spring 2012.

Table 7-6. Mean *Temperament in Middle Childhood Questionnaire* attentional focus and inhibitory control scores, by child characteristics, spring second grade: Spring 2013

	Spring second grade					
	Attentional Focus (X6ATTMCQ)			Inhibitory Control (X6INTMCQ)		
Characteristic	Number	Mean	Standard deviation	Number	Mean	Standard deviation
Total sample	12,660	3.47	1.122	12,660	3.67	0.845
Sex						
Male	6,440	3.20	1.131	6,440	3.42	0.848
Female	6,210	3.76	1.040	6,210	3.93	0.761
Race/ethnicity						
White, non-Hispanic	6,190	3.50	1.112	6,190	3.70	0.850
Black, non-Hispanic	1,450	3.23	1.134	1,450	3.42	0.849
Hispanic	3,340	3.47	1.133	3,340	3.70	0.818
Asian, non-Hispanic	970	3.85	1.009	970	3.91	0.791
Hawaiian, Other Pacific Islander,				70	3.69	0.849
non-Hispanic	70	3.58	1.129			
American Indian/Alaska Native,				100	3.68	0.810
non-Hispanic	100	3.46	1.119			
More than one race, non-Hispanic	530	3.51	1.097	530	3.68	0.830
School type						
Public school	11,430	3.46	1.127	11,430	3.66	0.846
Private school	1,230	3.58	1.057	1,230	3.71	0.835

NOTE: Items contributing to these scales come from the *Children's Behavior Questionnaire* (Putnam and Rothbart 2006) and the *Temperament in Middle Childhood Questionnaire* (Simonds and Rothbart 2004). Estimates weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 5. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2013.

7.1.2 Teacher-Reported Social Skills

In both the fall and spring first- and second-grade 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)[4] and included in the self-administered child-level teacher questionnaire. The social skills battery includes some items taken verbatim from the SSRS, some items that are modifications of original SSRS items, and some items that measure the same kinds of skills and behaviors captured in the SSRS 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 1998–99 (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 (6 items),[5] 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 6 items), and internalizing problem behaviors (3 out of 4 items). Scores on rated items were summed and divided by the number of items rated to derive the scale score. 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 externalizing problem behaviors more often). Variable names for the fall and spring first- and second-grade teacher scale scores, descriptions, value ranges, weighted means, and standard deviations for these scales are shown in table 7-7. Data for the individual items contributing to each scale are not included in the data file because of copyright restrictions.

[4] The *Social Skills Rating System* is a copyrighted instrument (1990 NCS Pearson) and has been adapted with permission. The *Social Skills Rating System* was developed by Gresham and Elliott (1990).

[5] For children who were in first grade during the first-grade data collections (rounds 3 and 4) and for all children in the second-grade data collections (rounds 5 and 6), the externalizing problem behaviors composite (X3TCHEXT, X4TCHEXT, X5TCHEXT, X6TCHEXT) is based on six items. This is different from how the composite was created for the kindergarten rounds (rounds 1 and 2) and for children in kindergarten during the first-grade data collection (X1TCHEXT, X2TCHEXT, X4KTCHEXT). One additional item was included at the end of the "Social Skills" section of the questionnaire in first and second grade. The item asked about the child's tendency to talk at times when the child was not supposed to be talking. The item was added because it had been included in the first-grade round of the ECLS-K and was factored into the calculation of that study's first-grade composite score.

Table 7-7. Teacher-reported social skills scales variable names, descriptions, value ranges, weighted means, and standard deviations: School years 2011–12 and 2012–13

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X3TCHCON	X3 Teacher Report Self-Control	4,660	1–4	3.21	0.591
X3TCHPER	X3 Teacher Report Interpersonal Skills	4,720	1–4	3.14	0.613
X3TCHEXT	X3 Teacher Report Externalizing Problem Behaviors	4,960	1–4	1.67	0.590
X3TCHINT	X3 Teacher Report Internalizing Problem Behaviors	4,850	1–4	1.48	0.483
X4TCHCON	X4 Teacher Report Self-Control	13,200	1–4	3.21	0.621
X4TCHPER	X4 Teacher Report Interpersonal Skills	13,290	1–4	3.14	0.657
X4TCHEXT	X4 Teacher Report Externalizing Problem Behaviors	13,400	1–4	1.73	0.619
X4TCHINT	X4 Teacher Report Internalizing Problem Behaviors	13,300	1–4	1.55	0.508
X4KTCHCON	X4K Teacher Report Self-Control	420	1–4	3.09	0.616
X4KTCHPER	X4K Teacher Report Interpersonal Skills	420	1–4	3.04	0.671
X4KTCHEXT	X4K Teacher Report Externalizing Problem Behaviors	420	1–4	1.78	0.614
X4KTCHINT	X4K Teacher Report Internalizing Problem Behaviors	420	1–4	1.62	0.498
X5TCHCON	X5 Teacher Report Self-Control	4,170	1–4	3.23	0.614
X5TCHPER	X5 Teacher Report Interpersonal Skills	4,180	1–4	3.13	0.621
X5TCHEXT	X5 Teacher Report Externalizing Problem Behaviors	4,430	1–4	1.65	0.610
X5TCHINT	X5 Teacher Report Internalizing Problem Behaviors	4,340	1–4	1.50	0.522
X6TCHCON	X6 Teacher Report Self-Control	12,470	1–4	3.22	0.629
X6TCHPER	X6 Teacher Report Interpersonal Skills	12,520	1–4	3.12	0.664
X6TCHEXT	X6 Teacher Report Externalizing Problem Behaviors	12,660	1–4	1.72	0.625
X6TCHINT	X6 Teacher Report Internalizing Problem Behaviors	12,580	1–4	1.59	0.528

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (3 for the fall first-grade round, 4 for the spring first-grade round, 5 for the fall second-grade round, and 6 for the spring second-grade round). X3 estimates weighted by W3CF3P3T0. X4 estimates weighted by W4CS4P_2T0. X5 estimates weighted by W6CF6P_2A0. X6 estimates weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011, spring 2012, fall 2012, and spring 2013.

Table 7-8 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 7-8. Reliability estimates for the teacher-reported social skills scales: School years 2011–12 and 2012–13

Variable name	Description	Number of items	Reliability coefficient
X3TCHCON	X3 Teacher Report Self-Control	4	.79
X3TCHPER	X3 Teacher Report Interpersonal Skills	5	.85
X3TCHEXT	X3 Teacher Report Externalizing Problem Behaviors	6	.88
X3TCHINT	X3 Teacher Report Internalizing Problem Behaviors	4	.77
X4TCHCON	X4 Teacher Report Self-Control	4	.81
X4TCHPER	X4 Teacher Report Interpersonal Skills	5	.86
X4TCHEXT	X4 Teacher Report Externalizing Problem Behaviors	6	.86
X4TCHINT	X4 Teacher Report Internalizing Problem Behaviors	4	.76
X4KTCHCON	X4K Teacher Report Self-Control	4	.79
X4KTCHPER	X4K Teacher Report Interpersonal Skills	5	.88
X4KTCHEXT	X4K Teacher Report Externalizing Problem Behaviors	6	.87
X4KTCHINT	X4K Teacher Report Internalizing Problem Behaviors	4	.73
X5TCHCON	X5 Teacher Report Self-Control	4	.80
X5TCHPER	X5 Teacher Report Interpersonal Skills	5	.85
X5TCHEXT	X5 Teacher Report Externalizing Problem Behaviors	6	.88
X5TCHINT	X5 Teacher Report Internalizing Problem Behaviors	4	.78
X6TCHCON	X6 Teacher Report Self-Control	4	.81
X6TCHPER	X6 Teacher Report Interpersonal Skills	5	.86
X6TCHEXT	X6 Teacher Report Externalizing Problem Behaviors	6	.87
X6TCHINT	X6 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). The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (3 for the fall first-grade round, 4 for the spring first-grade round, 5 for the fall second-grade round, and 6 for the spring second-grade round).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), fall 2011, spring 2012, fall 2012, and spring 2013.

Within-round correlations among the four social skills scales (self-control, interpersonal skills, externalizing problem behaviors, and internalizing problem behaviors) are presented in table 7-9.[6] Patterns of within-round correlations are similar across rounds, and all correlations are statistically significant. Self-control is positively correlated with interpersonal skills, with correlations ranging from .77 to .81 across rounds 1-6. Externalizing and internalizing problem behaviors are correlated in the expected direction with other social skills measures. Externalizing problem behaviors is negatively related to self-control and interpersonal skills, indicating that children rated higher on externalizing behaviors had

[6] Though the focus of this report is the first- and second-grade data collections, correlations for the kindergarten rounds are also included in table 7-9 because this information was not provided in the *ECLS-K:2011 Kindergarten Psychometric Report* (Najarian et al. 2018).

lower ratings of self-control and interpersonal skills. The correlation between externalizing problem behaviors and self-control ranges from -.68 and -.74 across rounds 1-6; the correlation between externalizing problem behaviors and interpersonal skills ranges from -.57 to -.63 across rounds 1-6. Internalizing problem behaviors is negatively associated with self-control and interpersonal skills, ranging from -.26 to -.36 and from -.31 to -.38, respectively, across rounds 1-6. Externalizing problem behaviors is positively related to internalizing problem behaviors, with correlations ranging from .26 to .33 across rounds 1-6.

Within-round correlations between interpersonal skills and self-control were consistently the strongest within-round correlation. The self-control scale includes items on control of emotions and behavior. Because the teacher observes the child in the context of peers and the self-control items are asked about behaviors that occur in the context of peers, it is not surprising that the self-control items are related to items in the interpersonal skills scale. Within-round correlations between externalizing problem behaviors and self-control and between externalizing problem behaviors and interpersonal skills were also consistently high across rounds. Internalizing problem behaviors show a pattern of correlations similar to externalizing problem behaviors, but correlations were lower in magnitude.

Table 7-9. Within-round correlations of teacher-reported social skills: School years 2010–11, 2011–12, and 2012–13

Round	Self-control (X*TCHCON)	Interpersonal skills (X*TCHPER)	Externalizing behavior (X*TCHEXT)	Internalizing behavior (X*TCHINT)
Round 1: Fall kindergarten				
Self-control	1.00			
Interpersonal skills	.79	1.00		
Externalizing behavior	-.71	-.59	1.00	
Internalizing behavior	-.26	-.31	.26	1.00
Round 2: Spring kindergarten				
Self-control	1.00			
Interpersonal skills	.81	1.00		
Externalizing behavior	-.74	-.63	1.00	
Internalizing behavior	-.29	-.34	.30	1.00
Round 3: Fall first grade				
Self-control	1.00			
Interpersonal skills	.77	1.00		
Externalizing behavior	-.71	-.58	1.00	
Internalizing behavior	-.30	-.35	.26	1.00
Round 4: Spring first grade				
Self-control	1.00			
Interpersonal skills	.80	1.00		
Externalizing behavior	-.72	-.62	1.00	
Internalizing behavior	-.31	-.35	.30	1.00
Round 5: Fall second grade				
Self-control	1.00			
Interpersonal skills	.78	1.00		
Externalizing behavior	-.68	-.57	1.00	
Internalizing behavior	-.33	-.38	.31	1.00
Round 6: Spring second grade				
Self-control	1.00			
Interpersonal skills	.81	1.00		
Externalizing behavior	-.73	-.63	1.00	
Internalizing behavior	-.36	-.38	.33	1.00

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). All correlations p <.001. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

Mean scores for the teacher-reported social skills subscales are presented by data collection round and child characteristics in tables 7-10 through 7-13.

Table 7-10. Mean teacher-reported self-control score, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall first grade			Spring first grade			Fall second grade			Spring second grade		
	(X3TCHCON)			(X4TCHCON)			(X5TCHCON)			(X6TCHCON)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	4,660	3.21	0.591	13,200	3.21	0.621	4,170	3.23	0.614	12,470	3.22	0.629
Sex												
Male	2,460	3.11	0.616	6,680	3.11	0.640	2,180	3.11	0.624	6,340	3.11	0.652
Female	2,200	3.32	0.543	6,510	3.32	0.582	2,000	3.36	0.578	6,120	3.34	0.580
Race/ethnicity												
White, non-Hispanic	1,790	3.26	0.576	6,400	3.25	0.610	1,610	3.26	0.593	6,110	3.25	0.617
Black, non-Hispanic	490	3.05	0.662	1,580	3.01	0.651	410	3.05	0.743	1,440	3.01	0.683
Hispanic	1,720	3.21	0.583	3,350	3.22	0.610	1,610	3.25	0.571	3,270	3.25	0.603
Asian, non-Hispanic	340	3.37	0.548	1,070	3.34	0.604	300	3.42	0.528	950	3.37	0.586
Hawaiian, Other Pacific Islander, non-Hispanic	30	3.01	0.809	80	3.34	0.575	20	3.53	0.446	70	3.27	0.546
American Indian/Alaska Native, non-Hispanic	90	3.07	0.475	110	3.21	0.594	70	3.18	0.569	100	3.17	0.638
More than one race, non-Hispanic	190	3.18	0.546	590	3.20	0.618	160	3.25	0.604	530	3.21	0.640
School type												
Public school	4,270	3.21	0.587	11,800	3.21	0.622	3,870	3.23	0.612	11,270	3.22	0.631
Private school	380	3.25	0.629	1,390	3.23	0.615	300	3.25	0.632	1,200	3.23	0.609

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). Fall first-grade estimates (X3) weighted by W3CF3P3T0. Spring first-grade estimates (X4) weighted by W4CS4P_2T0. Fall second-grade estimates (X5) weighted by W6CF6P_2A0. Spring second-grade estimates (X6) weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. *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), fall 2011, spring 2012, fall 2012, and spring 2013.

Table 7-11. Mean teacher-reported interpersonal skills score, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall first grade (X3TCHPER)			Spring first grade (X4TCHPER)			Fall second grade (X5TCHPER)			Spring second grade (X6TCHPER)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	4,720	3.14	0.613	13,290	3.14	0.657	4,180	3.13	0.621	12,520	3.12	0.664
Sex												
Male	2,460	3.02	0.617	6,700	3.01	0.664	2,150	2.99	0.628	6,330	2.98	0.672
Female	2,260	3.26	0.585	6,580	3.28	0.622	2,030	3.27	0.582	6,170	3.26	0.623
Race/ethnicity												
White, non-Hispanic	1,830	3.18	0.586	6,450	3.18	0.657	1,620	3.15	0.600	6,130	3.15	0.657
Black, non-Hispanic	500	3.00	0.694	1,580	2.96	0.665	420	3.01	0.703	1,440	2.94	0.708
Hispanic	1,720	3.11	0.607	3,380	3.15	0.640	1,570	3.14	0.594	3,280	3.12	0.643
Asian, non-Hispanic	360	3.29	0.599	1,070	3.20	0.653	310	3.18	0.626	950	3.20	0.639
Hawaiian, Other Pacific Islander, non-Hispanic	20	3.00	0.825	90	3.22	0.663	20	3.29	0.456	70	3.22	0.621
American Indian/Alaska Native, non-Hispanic	90	3.11	0.558	110	3.22	0.609	80	3.15	0.678	100	3.07	0.676
More than one race, non-Hispanic	200	3.06	0.620	590	3.13	0.645	160	3.10	0.662	540	3.11	0.664
School type												
Public school	4,340	3.13	0.615	11,880	3.14	0.660	3,870	3.12	0.623	11,310	3.11	0.665
Private school	380	3.22	0.584	1,390	3.19	0.628	310	3.25	0.580	1,210	3.18	0.648

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). Fall first-grade estimates (X3) weighted by W3CF3P3T0. Spring first-grade estimates (X4) weighted by W4CS4P_2T0. Fall second-grade estimates (X5) weighted by W6CF6P_2A0. Spring second-grade estimates (X6) weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. *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), fall 2011, spring 2012, fall 2012, and spring 2013.

Table 7-12. Mean teacher-reported externalizing problem behaviors score, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall first grade (X3TCHEXT)			Spring first grade (X4TCHEXT)			Fall second grade (X5TCHEXT)			Spring second grade (X6TCHEXT)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	4,960	1.67	0.590	13,400	1.73	0.619	4,430	1.65	0.610	12,660	1.72	0.625
Sex												
Male	2,610	1.78	0.626	6,780	1.86	0.657	2,300	1.78	0.645	6,440	1.86	0.660
Female	2,350	1.55	0.523	6,600	1.60	0.544	2,120	1.51	0.536	6,200	1.58	0.551
Race/ethnicity												
White, non-Hispanic	1,900	1.64	0.591	6,480	1.72	0.613	1,690	1.65	0.590	6,170	1.71	0.615
Black, non-Hispanic	510	1.85	0.642	1,600	1.95	0.686	430	1.84	0.733	1,460	1.96	0.693
Hispanic	1,840	1.63	0.563	3,420	1.67	0.584	1,710	1.56	0.556	3,340	1.66	0.592
Asian, non-Hispanic	380	1.50	0.460	1,090	1.54	0.516	330	1.45	0.428	970	1.54	0.541
Hawaiian, Other Pacific Islander, non-Hispanic	30	1.71	0.705	90	1.66	0.553	20	1.52	0.554	70	1.61	0.580
American Indian/Alaska Native, non-Hispanic	100	1.77	0.518	120	1.78	0.574	80	1.72	0.627	100	1.72	0.593
More than one race, non-Hispanic	200	1.68	0.564	590	1.75	0.607	160	1.74	0.680	540	1.74	0.615
School type												
Public school	4,560	1.67	0.588	11,990	1.73	0.622	4,110	1.65	0.611	11,430	1.73	0.629
Private school	400	1.60	0.607	1,390	1.72	0.582	320	1.66	0.591	1,230	1.71	0.582

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). Fall first-grade estimates (X3) weighted by W3CF3P3T0. Spring first-grade estimates (X4) weighted by W4CS4P_2T0. Fall second-grade estimates (X5) weighted by W6CF6P_2A0. Spring second-grade estimates (X6) weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding.*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), fall 2011, spring 2012, fall 2012, and spring 2013.

Table 7-13. Mean teacher-reported internalizing problem behaviors score, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall first grade (X3TCHINT)			Spring first grade (X4TCHINT)			Fall second grade (X5TCHINT)			Spring second grade (X6TCHINT)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	4,850	1.48	0.483	13,310	1.55	0.508	4,340	1.50	0.522	12,580	1.593	0.528
Sex												
Male	2,530	1.51	0.499	6,720	1.57	0.518	2,260	1.54	0.559	6,380	1.62	0.544
Female	2,310	1.46	0.464	6,570	1.53	0.497	2,080	1.45	0.475	6,180	1.57	0.510
Race/ethnicity												
White, non-Hispanic	1,870	1.47	0.460	6,470	1.55	0.498	1,680	1.49	0.508	6,160	1.60	0.524
Black, non-Hispanic	500	1.52	0.499	1,580	1.61	0.539	420	1.58	0.597	1,440	1.64	0.560
Hispanic	1,790	1.49	0.499	3,380	1.54	0.516	1,650	1.48	0.496	3,310	1.57	0.527
Asian, non-Hispanic	360	1.38	0.460	1,070	1.42	0.423	320	1.36	0.397	960	1.51	0.476
Hawaiian, Other Pacific Islander, non-Hispanic	30	1.70	0.566	80	1.44	0.441	20	1.28	0.362	70	1.48	0.464
American Indian/ Alaska Native, non-Hispanic	100	1.57	0.597	120	1.68	0.567	80	1.57	0.592	100	1.66	0.578
More than one race, non-Hispanic	190	1.52	0.508	590	1.55	0.517	170	1.60	0.597	530	1.58	0.498
School type												
Public school	4,460	1.48	0.486	11,900	1.55	0.512	4,020	1.50	0.524	11,350	1.60	0.531
Private school	390	1.50	0.449	1,390	1.56	0.470	320	1.51	0.492	1,220	1.56	0.490

NOTE: Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). Fall first-grade estimates (X3) weighted by W3CF3P3T0. Spring first-grade estimates (X4) weighted by W4CS4P_2T0. Fall second-grade estimates (X5) weighted by W6CF6P_2A0. Spring second-grade estimates (X6) weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding.*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), fall 2011, spring 2012, fall 2012, and spring 2013.

7.1.3 Teacher-Reported Approaches to Learning Items and Scale

The fall and spring first- and second-grade child-level teacher questionnaires 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).[7] These items were presented in the same set of items as the 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. Scores on rated items were summed and divided by the number of items rated to derive the scale score. 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 and spring first- and second-grade teacher-reported Approaches to Learning scale scores are shown in table 7-14. 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.

Table 7-14. Teacher-reported Approaches to Learning scale variable names, descriptions, value ranges, weighted means, and standard deviations: School years 2011–12 and 2012–13

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X3TCHAPP	X3 Teacher Report Approaches to Learning	5,020	1–4	3.04	0.677
X4TCHAPP	X4 Teacher Report Approaches to Learning	13,450	1–4	3.07	0.700
X4KTCHAPP	X4K Teacher Report Approaches to Learning	420	1–4	2.94	0.704
X5TCHAPP	X5 Teacher Report Approaches to Learning	4,510	1–4	3.05	0.688
X6TCHAPP	X6 Teacher Report Approaches to Learning	12,690	1–4	3.07	0.707

NOTE: The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (3 for the fall first-grade round, 4 for the spring first-grade round, 5 for the fall second-grade round, and 6 for the spring second-grade round). X3 estimates weighted by W3CF3P3T0. X4 estimates weighted by W4CS4P_2T0. X5 estimates weighted by W6CF6P_2A0. X6 estimates weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2011, spring 2012, fall 2012, and spring 2013.

[7] 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 the ECLS-K. The seventh item (follows classroom rules) was added in the first-grade round of the ECLS-K.

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 first-grade child-level teacher questionnaire begin with “T3.” Those for the item-level data from the spring first-grade child-level teacher questionnaire for children in first grade begin with “T4,” while those for children held back in kindergarten begin with “T4K.” Variable names for the fall of second grade begin with “T5,” and those for the spring of second grade begin with “T6.”

Within-round correlations between teacher-reported approaches to learning and teacher-reported social skills were examined for all rounds of data collection (kindergarten through second grade) and are presented in table 7-15.[8] All within-round correlations were significant across all rounds. Approaches to learning was correlated with interpersonal skills and self-control within rounds, correlations ranging from .71 to .74 for interpersonal skills and from .68 to .72 for self-control. There was a significant negative correlation between approaches to learning and externalizing problem behaviors and between approaches to learning and internalizing problem behaviors. Across all rounds, externalizing problem behaviors was associated with approaches to learning (range -.58 to -.62), and internalizing problem behaviors was also associated with approaches to learning (range -.30 to -.42).

Table 7-15. Within-round correlations of the teacher-reported Approaches to Learning scale and the teacher-reported social skills scales: School years 2010–11, 2011–12, and 2012–13

Round	Self-control (X*TCHCON)	Interpersonal skills (X*TCHPER)	Externalizing behavior (X*TCHEXT)	Internalizing behavior (X*TCHINT)
Approaches to Learning				
Round 1: X1TCHAPP	.71	.74	-.59	-.30
Round 2: X2TCHAPP	.72	.73	-.61	-.35
Round 3: X3TCHAPP	.68	.71	-.60	-.37
Round 4: X4TCHAPP	.70	.72	-.61	-.37
Round 5: X5TCHAPP	.68	.72	-.58	-.40
Round 6: X6TCHAPP	.70	.73	-.62	-.42

NOTE: The “*” in the variables names is to be substituted for the data collection round number (1 to 6). All correlations $p < .001$. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

Mean scores for the teacher-reported Approaches to Learning scale are presented by data collection round and child characteristics in table 7-16.

[8] Although the focus of this report is the first- and second-grade data collections, correlations for the kindergarten rounds are also included in table 7-15 because this information was not included in the *ECLS-K:2011 Kindergarten Psychometric Report* (Najarian et al. 2018).

Table 7-16. Mean teacher-reported Approaches to Learning scale scores, by data collection round and child characteristics: School years 2011–12 and 2012–13

	Fall first grade (X3TCHAPP)			Spring first grade (X4TCHAPP)			Fall second grade (X5TCHAPP)			Spring second grade (X6TCHAPP)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	5,020	3.04	0.677	13,450	3.07	0.700	4,510	3.05	0.688	12,690	3.07	0.707
Sex												
Male	2,640	2.90	0.673	6,800	2.91	0.702	2,350	2.88	0.693	6,450	2.89	0.717
Female	2,380	3.20	0.645	6,630	3.24	0.656	2,160	3.22	0.636	6,230	3.25	0.647
Race/ethnicity												
White, non-Hispanic	1,920	3.09	0.658	6,490	3.11	0.687	1,720	3.07	0.665	6,190	3.10	0.693
Black, non-Hispanic	520	2.87	0.731	1,600	2.86	0.714	440	2.89	0.768	1,460	2.86	0.736
Hispanic	1,860	3.04	0.674	3,440	3.06	0.706	1,740	3.04	0.681	3,350	3.06	0.711
Asian, non-Hispanic	390	3.27	0.617	1,100	3.29	0.653	330	3.29	0.584	970	3.30	0.634
Hawaiian, Other Pacific Islander, non-Hispanic	30	2.79	0.580	90	3.13	0.677	20	3.00	0.634	70	3.18	0.627
American Indian/Alaska Native, non-Hispanic	100	2.92	0.688	120	2.98	0.682	80	3.08	0.743	100	3.09	0.722
More than one race, non-Hispanic	210	2.95	0.661	590	3.09	0.680	170	2.99	0.712	540	3.08	0.692
School type												
Public school	4,610	3.04	0.681	12,030	3.06	0.705	4,180	3.03	0.693	11,460	3.06	0.711
Private school	410	3.14	0.625	1,400	3.13	0.642	330	3.20	0.613	1,230	3.14	0.664

NOTE: Fall first-grade estimates (X3) weighted by W3CF3P3T0. Spring first-grade estimates (X4) weighted by W4CS4P_2T0. Fall second-grade estimates (X5) weighted by W6CF6P_2A0. Spring second-grade estimates (X6) weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. *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), fall 2011, spring 2012, fall 2012, and spring 2013.

7.1.4 *Student-Teacher Relationship Scale*

The *Student-Teacher Relationship Scale (STRS)* (Pianta and Steinberg 1992) is a 15-item, teacher-reported measure of closeness and conflict between the teacher and child. As part of the spring first-grade and spring second-grade child-level teacher questionnaires, 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 authors: closeness and conflict. The closeness scale score is the average rating on the seven items included in the

Closeness scale of the *STRS*, while the conflict scale score is the average rating on the eight items included in Conflict scale of the *STRS*. A score was computed when the respondent provided a rating on at least five of the seven or eight items that composed the scales. Scores on rated items were summed and divided by the number of items rated to derive the scale score. 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. Higher scores on the closeness scale indicate that the teacher perceived he or she had a closer relationship with the child. Higher 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 7-17.

Table 7-17. *Student-Teacher Relationship Scale* variable names, descriptions, value ranges, weighted means, and standard deviations: Spring 2012 and spring 2013

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X4CLSNSS	X4 Teacher Report Closeness	13,420	1–5	4.30	0.662
X4CNFLCT	X4 Teacher Report Conflict	13,420	1–5	1.64	0.792
X4KCLSNSS	X4K Teacher Report Closeness	420	1–5	4.27	0.693
X4KCNFLCT	X4K Teacher Report Conflict	420	1–5	1.82	0.875
X6CLSNSS	X6 Teacher Report Closeness	12,680	1–5	4.24	0.687
X6CNFLCT	X6 Teacher Report Conflict	12,680	1–5	1.63	0.794

NOTE: Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Steinberg 1992). The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (4 for the spring first-grade round and 6 for the spring second-grade round). X4 estimates weighted by W4CS4P_2T0. X6 estimates weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012 and spring 2013.

Table 7-18 presents the internal consistency reliability coefficients (Cronbach's alpha) for the teacher-reported *STRS* closeness and conflict scores for first and second grades. In the spring of first grade and the spring of second grade, the closeness scale (X4CLSNSS, X4KCLSNSS, and X6CLSNSS) has a reliability estimate that ranges from .86 to .87, and the conflict scale (X4CNFLCT, X4KCNFLCT, and X6CNFLCT) has an internal consistency reliability estimate that ranges from .88 to .90, as measured by Cronbach's alpha. The closeness and conflict scores are negatively correlated within rounds, with correlations ranging from -.34 to -.51. Recently, the study received copyright permission to include item-level data from the *STRS* on the ECLS-K:2011 restricted-use data files. While item-level data from the *STRS* are not on the kindergarten through first-grade data file, they are included on the kindergarten through second-grade restricted-use data file.

Table 7-18. Reliability estimates for the *Student-Teacher Relationship Scale* teacher-reported closeness and conflict scores: Spring 2012 and spring 2013

Variable name	Description	Number of items	Reliability coefficient
X4CLSNSS	X4 Teacher Report Closeness	7	.86
X4CNFLCT	X4 Teacher Report Conflict	8	.89
X4KCLSNSS	X4 Teacher Report Closeness	7	.87
X4KCNFLCT	X4 Teacher Report Conflict	8	.88
X6CLSNSS	X6 Teacher Report Closeness	7	.87
X6CNFLCT	X6 Teacher Report Conflict	8	.90

NOTE: Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Steinberg 1992). The name and description for each variable in the table begin with an "X," indicating that it is a derived/calculated variable, and a data collection round number (4 for the spring first-grade round and 6 for the spring second-grade round).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2012 and spring 2013.

Mean scores for the *Student-Teacher Relationship Scale* are presented by data collection round and child characteristics in tables 7-19 and 7-20.

Table 7-19. Mean *Student-Teacher Relationship Scale* teacher-reported closeness score, by data collection round and child characteristics: Spring 2012 and spring 2013

	Spring first grade (X4CLSNSS)			Spring second grade (X6CLSNSS)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	13,420	4.30	0.662	12,680	4.24	0.687
Sex						
Male	6,790	4.19	0.690	6,450	4.10	0.715
Female	6,610	4.42	0.608	6,220	4.39	0.622
Race/ethnicity						
White, non-Hispanic	6,490	4.37	0.634	6,200	4.32	0.651
Black, non-Hispanic	1,600	4.25	0.643	1,450	4.17	0.715
Hispanic	3,430	4.21	0.708	3,340	4.13	0.707
Asian, non-Hispanic	1,090	4.19	0.695	960	4.13	0.733
Hawaiian, Other Pacific Islander, non-Hispanic	90	4.31	0.615	70	4.15	0.690
American Indian/Alaska Native, non-Hispanic	120	4.25	0.598	100	4.18	0.821
More than one race, non-Hispanic	590	4.29	0.654	540	4.21	0.674
School type						
Public school	12,010	4.29	0.667	11,450	4.23	0.689
Private school	1,400	4.42	0.598	1,230	4.33	0.653

NOTE: Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Steinberg 1992). Spring first-grade estimates (X4) weighted by W4CS4P_2T0. Spring second grade estimates (X6) weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 5. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. *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), spring 2012 and spring 2013.

Table 7-20. Mean *Student-Teacher Relationship Scale* teacher-reported conflict score, by data collection round and child characteristics: Spring 2012 and spring 2013

	Spring first grade (X4CNFLCT)			Spring second grade (X6CNFLCT)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	13,420	1.64	0.792	12,680	1.63	0.794
Sex						
Male	6,790	1.79	0.864	6,450	1.79	0.865
Female	6,620	1.48	0.673	6,220	1.46	0.671
Race/ethnicity						
White, non-Hispanic	6,490	1.61	0.777	6,210	1.59	0.759
Black, non-Hispanic	1,600	1.93	0.956	1,450	1.96	0.989
Hispanic	3,430	1.57	0.712	3,340	1.57	0.722
Asian, non-Hispanic	1,090	1.45	0.632	970	1.45	0.637
Hawaiian, Other Pacific Islander, non-Hispanic	90	1.60	0.754	70	1.52	0.640
American Indian/Alaska Native, non-Hispanic	120	1.69	0.734	100	1.65	0.761
More than one race, non-Hispanic	590	1.68	0.777	540	1.66	0.846
School type						
Public school	12,010	1.64	0.792	11,450	1.63	0.796
Private school	1,400	1.62	0.791	1,230	1.62	0.778

NOTE: Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Steinberg 1992). Spring first-grade estimates (X4) weighted by W4CS4P_2T0. Spring second grade estimates (X6) weighted by W6CS6P_2T0. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 5. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. *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), spring 2012 and spring 2013.

7.2 Parent Measures

7.2.1 Parent-Reported Social Skills

In the spring first-grade 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. Like the teacher-reported items, these parent items also are based on items from the *Social Skills Rating System*. The social skills items were not included in the spring second-grade parent interview.

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 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 items), sad/lonely (3 out of 4 items), and impulsive/overactive (2 out of 2 items). Scores on rated items were summed and divided by the number of items rated to derive the scale score. 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 spring first-grade parent scores are shown in table 7-21. Data for the individual items contributing to each scale are not included in the data file because of copyright restrictions.

Table 7-21. Parent-reported social skills scales variable names, descriptions, value ranges, weighted means, and standard deviations: Spring 2012

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X4PRNCON	X4 Parent Report Self-Control	12,560	1–4	3.02	0.495
X4PRNSOC	X4 Parent Report Social Interaction	12,590	1–4	3.45	0.544
X4PRNSAD	X4 Parent Report Sad/Lonely	12,540	1–4	1.46	0.386
X4PRNIMP	X4 Parent Report Impulsive/Overactive	12,460	1–4	1.88	0.664

NOTE: Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). Estimates weighted byW4CS4P_40. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

Table 7-22 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 7-22. Reliability estimates for the parent-reported social skills scales: Spring 2012

Variable name	Description	Number of items	Reliability coefficient
X4PRNCON	X4 Parent Report Self-Control	5	.73
X4PRNSOC	X4 Parent Report Social Interaction	3	.69
X4PRNSAD	X4 Parent Report Sad/Lonely	4	.62

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), spring 2012.

Correlations among the four parent-reported social skills scales (self-control, social interaction, sad/lonely, and impulsive/overactive) in the spring of first grade are presented in table 7-23. All correlations are significant. Self-control and social interaction are positively related to one another and are negatively related to sad/lonely and impulsive/overactive behaviors. Sad/lonely and impulsive/overactive behaviors were positively related to one another.

Table 7-23. Correlations of parent-reported social skills in the spring of first grade: Spring 2012

Social skills	Self-control (X4PRNCON)	Social interaction (X4PRNSOC)	Sad/lonely (X4PRNSAD)	Impulsive/ overactive (X4PRNIMP)
Self-control	1.00			
Social interaction	.23	1.00		
Sad/lonely	-.36	-.24	1.00	
Impulsive/overactive	-.45	-.09	.30	1.00

NOTE: Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). All correlations $p < .001$. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

Correlations among parent-reported and teacher-reported social skills in the spring of first grade are presented in table 7-24. Parent-reported social skills and teacher-reported social skills were not highly correlated, with the magnitude of the correlations ranging from .07 to .30. This may be because although there are similarities in how some of the scales are labeled, the items making up the scales for teachers and parents are different. Further, parents and teachers spend time with children in different settings, and the child's behaviors may be different depending on the context. Parent-reported self-control and social interaction were positively correlated with teacher-reported self-control and interpersonal skills, and they were negatively correlated with teacher-reported externalizing and internalizing problem behaviors. Parent-reported sad/lonely and impulsive/overactive behaviors were positively correlated with teacher-reported internalizing and externalizing behaviors and negatively correlated with teacher-reported self-control and interpersonal skills.

Table 7-24. Correlations between teacher-reported social skills and parent-reported social skills in the spring of first grade: Spring 2012

	Teacher report			
Parent report	Self-control (X4TCHCON)	Interpersonal skills (X4TCHPER)	Internalizing behavior (X4TCHINT	Externalizing behavior (X4TCHEXT)
Self-control (X4PRNCON)	.21	.21	-.10	-.22
Social interaction (X4PRNSOC)	.13	.19	-.11	-.07
Sad/lonely (X4PRNSAD)	-.10	-.11	.16	.10
Impulsive/overactive (X4PRNIMP)	-.25	-.22	.09	.30

NOTE: Items contributing to the teacher-reported and parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). All correlations $p < .001$. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

Mean scores for the parent-reported social skills subscales from the spring first-grade data collection are presented by selected child characteristics in tables 7-25 through 7-28.

Table 7-25. Mean parent-reported self-control score, by child characteristics: Spring 2012

	Spring first grade (X4PRNCON)		
Characteristic	Number	Mean	Standard deviation
Total sample	12,560	3.02	0.495
Sex			
Male	6,400	2.98	0.499
Female	6,160	3.05	0.487
Race/ethnicity			
White, non-Hispanic	6,280	3.03	0.464
Black, non-Hispanic	1,330	3.01	0.575
Hispanic	3,170	2.98	0.520
Asian, non-Hispanic	1,010	3.06	0.447
Hawaiian, Other Pacific Islander, non-Hispanic	70	3.19	0.373
American Indian/Alaska Native, non-Hispanic	90	2.99	0.564
More than one race, non-Hispanic	600	3.05	0.457
School type			
Public school	10,930	3.01	0.498
Private school	1,270	3.10	0.441

NOTE: Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). Estimates weighted by W4CS4P_40. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

Table 7-26. Mean parent-reported social interaction score, by child characteristics: Spring 2012

	Spring first grade (X4PRNSOC)		
Characteristic	Number	Mean	Standard deviation
Total sample	12,590	3.45	0.544
Sex			
Male	6,420	3.40	0.564
Female	6,170	3.50	0.516
Race/ethnicity			
White, non-Hispanic	6,290	3.52	0.495
Black, non-Hispanic	1,330	3.42	0.562
Hispanic	3,180	3.32	0.590
Asian, non-Hispanic	1,020	3.26	0.606
Hawaiian, Other Pacific Islander, non-Hispanic	70	3.46	0.562
American Indian/Alaska Native, non-Hispanic	90	3.45	0.544
More than one race, non-Hispanic	600	3.55	0.513
School type			
Public school	10,950	3.44	0.548
Private school	1,280	3.52	0.481

NOTE: Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). Estimates weighted by W4CS4P_40. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

Table 7-27. Mean parent-reported sad/lonely score, by child characteristics: Spring 2012

	Spring first grade (X4PRNSAD)		
Characteristic	Number	Mean	Standard deviation
Total sample	12,540	1.46	0.386
Sex			
Male	6,400	1.46	0.387
Female	6,150	1.47	0.385
Race/ethnicity			
White, non-Hispanic	6,280	1.47	0.362
Black, non-Hispanic	1,320	1.48	0.427
Hispanic	3,170	1.41	0.403
Asian, non-Hispanic	1,010	1.54	0.413
Hawaiian, Other Pacific Islander, non-Hispanic	70	1.56	0.459
American Indian/Alaska Native, non-Hispanic	90	1.54	0.366
More than one race, non-Hispanic	600	1.48	0.368
School type			
Public school	10,910	1.46	0.390
Private school	1,270	1.46	0.347

NOTE: Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). Estimates weighted by W4CS4P_40. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

Table 7-28. Mean parent-reported impulsive/overactive score, by child characteristics: Spring 2012

	Spring first grade (X4PRNIMP)		
Characteristic	Number	Mean	Standard deviation
Total sample	12,460	1.88	0.664
Sex			
Male	6,360	1.97	0.687
Female	6,090	1.79	0.625
Race/ethnicity			
White, non-Hispanic	6,260	1.89	0.650
Black, non-Hispanic	1,310	1.98	0.727
Hispanic	3,150	1.82	0.662
Asian, non-Hispanic	980	1.81	0.630
Hawaiian, Other Pacific Islander, non-Hispanic	70	1.89	0.568
American Indian/Alaska Native, non-Hispanic	90	1.96	0.665
More than one race, non-Hispanic	600	1.87	0.652
School type			
Public school	10,850	1.89	0.669
Private school	1,270	1.80	0.597

NOTE: Items contributing to the parent-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). Estimates weighted by W4CS4P_40. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

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).[9] These items were asked in section SSQ (Social Skills, Problem Behaviors, and Approaches to Learning) of the parent interview within the same set of items as the social skills items based on the *Social Skills Rating System* (described in section 7.2.1), 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

[9] 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.

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. Scores on rated items were summed and divided by the number of items rated to derive the scale score. Higher scale scores indicate that the child exhibited positive learning behaviors more often. The variable name, description, value range, weighted mean, and standard deviation for the spring first-grade parent Approaches to Learning scale score is shown in table 7-29. The Approaches to Learning scale has an internal consistency reliability estimate of .74 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 spring first-grade parent interview begin with "P4."

Parent-reported approaches to learning and parent-reported social skills are significantly related. More specifically, parent-reported approaches to learning was positively correlated with parent-reported self-control (r = .26, p < .001) and social interaction (r = .50, p < .001), while it was negatively correlated with sad/lonely behaviors (r = -.18, p < .001) and impulsive/overactive behaviors (r = -.16, p < .001).

The correlation between teacher-reported approaches to learning and parent-reported approaches to learning in spring first-grade is significant yet somewhat low (r = .27, p < .001). Parents and teachers spend time with children in different settings and, therefore, have different kinds of experiences with children. In addition, although there is some overlap of items on the parent and teacher versions of the approaches to learning scales, different items comprise the teacher and parent approaches to learning scales.

Table 7-29. Parent-reported Approaches to Learning scale variable names, descriptions, value ranges, weighted means, and standard deviations: Spring 2012

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X4PRNAPP	X4 Parent Report Approaches to Learning	12,550	1–4	3.08	0.497

NOTE: Estimates weighted by W4CS4P_40. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

Mean scores for the parent-reported Approaches to Learning scale are presented by child characteristics for the spring first-grade data collection in table 7-30.

Table 7-30. Mean parent-reported Approaches to Learning scale scores, by child characteristics: Spring 2012

	Spring first grade (X4PRNAPP)		
Characteristic	Number	Mean	Standard deviation
Total sample	12,550	3.08	0.497
Sex			
Male	6,400	3.02	0.508
Female	6,150	3.16	0.475
Race/ethnicity			
White, non-Hispanic	6,280	3.13	0.470
Black, non-Hispanic	1,330	3.09	0.535
Hispanic	3,170	2.99	0.521
Asian, non-Hispanic	1,010	3.00	0.516
Hawaiian, Other Pacific Islander, non-Hispanic	70	3.08	0.418
American Indian/Alaska Native, non-Hispanic	90	3.09	0.435
More than one race, non-Hispanic	600	3.17	0.474
School type			
Public school	10,930	3.08	0.500
Private school	1,280	3.15	0.467

NOTE: Estimates weighted by W4CS4P_40. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. The unweighted *n* is the rounded number of cases with a valid score. The range of possible values is 1 to 4. Detail may not sum to total due to rounding and/or missing data. Estimates produced from the Electronic Codebook may differ slightly from estimates shown due to rounding. Estimates produced from the Electronic Codebook may differ slightly from estimates shown 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 2012.

REFERENCES

Adler, C. R. (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.

Beguin, A., and Hansen, B. (2001). *Effect of Noncompensatory Multidimensionality on Separate and Concurrent Estimation in IRT Observed Score Equating*. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Bock, R. D., Thissen, D., and Zimowski M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*: 197–211.

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.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika 16*(3): 297–334.

Davidson, M.C., Amso, D., Cruess-Anderson, L., and Diamond, A. (2006). Development of Cognitive Control and Executive Function from 4-13 Years: Evidence from Manipulations of Memory, Inhibition and Task Switching. *Neuropsychologia*, *44*: 2037–2078.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1): 1-38.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology, 64*: 135–168.

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

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.

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

Ferguson, G.A.,(1941) The Factorial Interpretation of Test Difficulty. *Psychometrika, 6*(5): 323-329.

Gresham, F.M., and Elliott, S.N. (1990). *Social Skills Rating System.* Circle Pines, MN: American Guidance Service.

Hansen, B., and Beguin, A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common Item Equating Design. *Applied Psychological Measurement*, *26*(1): 3–24.

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, DE: 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 America: 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). (1968). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Publishers.

Lord, F.M., and Novick, M.R., with contributions by Allen Birnbaum. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addision-Wesley.

McGrew, K. S., and Woodcock, R. W. (2001). Technical Manual. *Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.

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.

Meng, H. (2007). *A Comparison of IRT Calibration Methods for Mixed-Format Tests in Vertical Scaling.* Dissertation, University of Iowa.

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. (1991). PARSCALE: *Parameter scaling of rating data* [computer program]. Chicago: Scientific Software, Inc.

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.

National Assessment Governing Board. (n.d.). *Reading Framework for the 1992 and 1994 National Assessment of Educational Progress.* NAEP Reading Consensus Project. Washington, DC: U.S. Department of Education.

National Assessment Governing Board. (1995). *Mathematics Framework for the 1996 National Assessment of Educational Progress.* Washington, DC: Government Printing Office.

National Assessment Governing Board. (2004a). *Mathematics Framework for the 2005 National Assessment of Educational Progress.* Washington, DC: U.S. Government Printing Office. Retrieved 11/10/2015 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 11/10///2015 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.

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

Pianta, R. C., and Steinberg, M. S. (1992). Teacher–Child Relationships and the Process of Adjusting to School. In R. C. Pianta (Ed.), *Beyond the Parent: The Role of Other Adults in Children's Lives* (pp. 61-80). San Francisco: Jossey-Bass.

Pollack, J.M., Rock, D.A., Weiss, M.J., Atkins-Burnett, S., Tourangeau, K., West, J., and Hausken, E.G. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

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.

Reise, S.P., Horan, W. P., and Blanchard, J J. (2011). The Challenges of Fitting an Item Response Theory Model to the Social Anhedonia Scale. *Journal of Personality Assessment, 93*(3): 213–224.

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.

Rothbart, M.K., Ahadi, S.A., Hershey, K., and Fisher, P. (2001). Investigations of Temperament at Three to Seven Years: The Children's Behavior Questionnaire. *Child Development, 72*(5): 1394–1408.

Simonds, J., and Rothbart, M.K. (2004, October). *The Temperament in Middle Childhood Questionnaire (TMCQ): A Computerized Self-Report Measure of Temperament for Ages 7-10*. Poster session presented at the Occasional Temperament Conference, Athens, GA.

Slotkin, J., Kallen, M., Griffith, J., Magasi, S., Salsman, J., Nowinski, C., and Gershon, R. (2012). *NIH Toolbox Technical Manual.* © 2006–12 National Institutes of Health and Northwestern University. Retrieved January 4, 2016 from http://www.nihtoolbox.org/HowDoI/TechnicalManual/Pages/default.aspx.

Slotkin, J., Nowinski, C., Hays, R., Beaumont, J., Griffith, J., Magasi, S., Salsman, J., and Gershon, R. (2012). *NIH Toolbox Scoring and Interpretation Guide.* © 2006–12 National Institutes of Health and Northwestern University. Retrieved January 4, 2016 from http://www.nihtoolbox.org/HowDoI/Pages/ScoringAndInterpretation.aspx.

Stocking, M.L., and Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A.G., Hagedorn, M.C., Daly, P., and Najarian, M. (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.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Hagedorn, M.C., Leggitt, J., and Najarian, M. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–First Grade Data File and Electronic Codebook, Public Version* (NCES 2015-078). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Tsai, T., Hanson, B.A., Kolen, M.J., and Forsyth, R.A. (2001). A Comparison of Bootstrap Standard Errors of IRT Equating Methods for the Common Item Non-Equivalent Group Design. *Applied Measurement in Education, 14*(1): 17–30.

U.S. Department of Education, National Center for Education Statistics. (2013). *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) Restricted-Use Kindergarten Data File and Electronic Codebook* (NCES 2013-060).

U.S. Department of Education, National Center for Education Statistics. (2014). *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) Restricted-Use Kindergarten–First Grade Data File and Electronic Codebook* (NCES 2015-070).

U.S. Department of Education, National Center for Education Statistics. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Public-Use Kindergarten–First Grade Data File and Electronic Codebook* (NCES 2015-086).

U.S. Department of Education, National Center for Education Statistics. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Restricted-Use Kindergarten–Second Grade Data File and Electronic Codebook* (NCES 2015-050).

U.S. Department of Education, National Center for Education Statistics. (2017). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Public-Use Kindergarten–Second Grade Data File and Electronic Codebook* (NCES 2017-286).

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.

Weiner, I.B., Schinka, J.A., and Velicer, W.F. (Eds.). (2004). *Handbook of Psychology, Vol. 2: Research Methods in Psychology*. New York: John Wiley & Sons.

Weintraub, S., Zelazo, P.D., Bauer, P.J., Wallner-Allen, K., Dikmen, S.S., Heaton, R.K., Tulsky, D.S., Slotkin, J., Blitz, D., Carlozzi, N.E., Havlik, R., Beaumont, J.L., Mungas, D., Manly, J.J., Nowinski, C., Borosh, B., and Gershon, R. (2013). NIH Toolbox Cognitive Function Battery (CFB): Introduction and Pediatric Data. In P.D. Zelazo and P.J. Bauer (Eds.), *National Institutes of Health Toolbox—Cognitive Function Battery: Validation for Children between 3 and 15 years.* Monographs of the Society for Research in Child Development, 78 (4, Serial No. 309).

Woodcock, R. W., McGrew, K. S., and Mather, N. (2001). *Woodcock-Johnson III Tests of Cognitive Ablities.* Itasca, IL: Riverside Publishing.

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

Yen, W.M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, *8*(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.

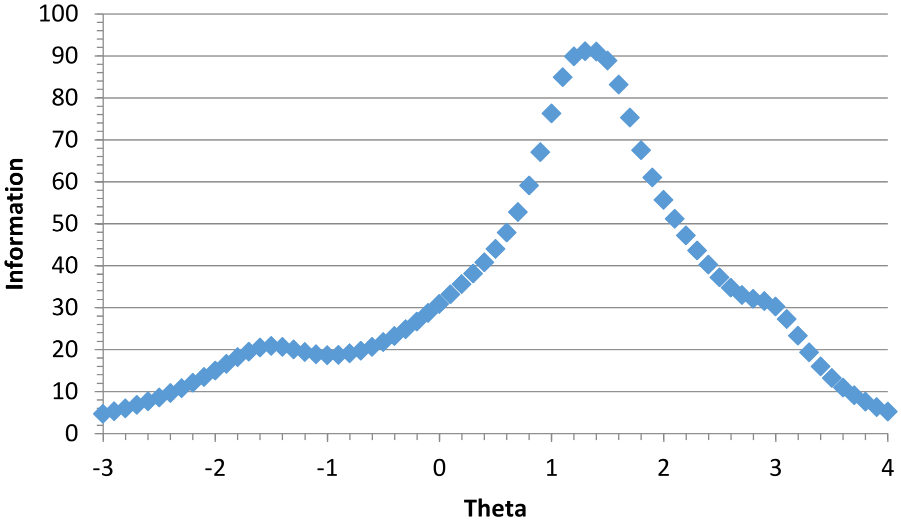
Zelazo, P.D., Anderson, J.E., Richler, J., Wallner-Allen, K., Beaumont, J.L., and Weintraub, S. (2013). NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. In P.D. Zelazo and P.J. Bauer (Eds.), *National Institutes of Health Toolbox Cognition Battery (NIH Toolbox CB): Validation for Children Between 3 and 15 Years. Monographs of the Society for Research in Child Development*.

This page intentionally left blank.

APPENDIX A

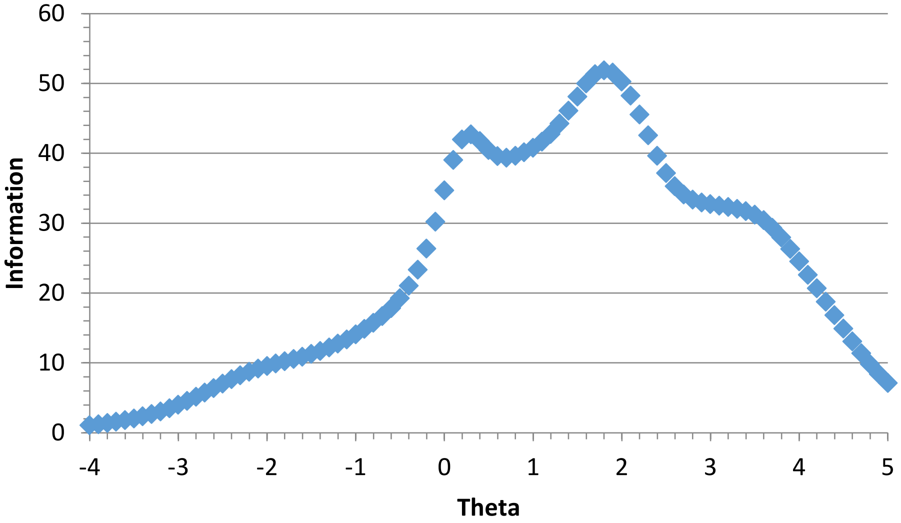
PLOTS OF IRT TEST INFORMATION FUNCTIONS FOR READING, MATHEMATICS, AND SCIENCE

Figure A-1. IRT test information function for reading for school years 2010–11, 2011–12, and 2012-13



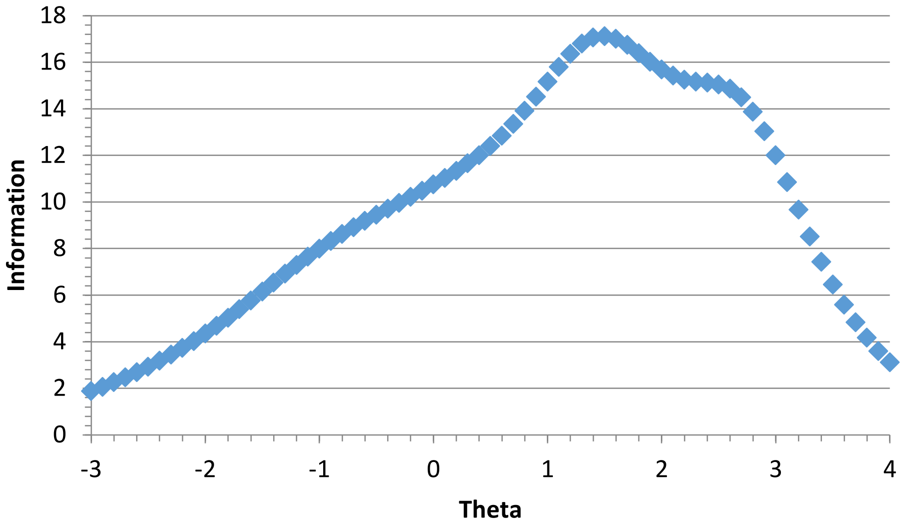
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2010, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

Figure A-2. IRT test information function for mathematics for school years 2010–11, 2011–12, and 2012–13



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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

Figure A-3. IRT test information function for science for school years 2010–11, 2011–12, and 2012–13



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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

APPENDIX B
IRT ESTIMATION USING PARSCALE

This appendix provides more detail on how the raw item responses were prepared for use in PARSCALE, how PARSCALE estimates the IRT model parameters, and what quality control checks were 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 kindergarten, spring kindergarten, fall first grade, spring first grade, fall second grade and spring second grade. 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 fall and spring, within-grade) were stacked, and frequency counts were checked to confirm the accuracy of the concatenated files, by grade. If applicable, 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.

[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.

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

$$\begin{aligned} L(\beta) &= \prod_g \prod_{j:g} \int_\theta P(x_{j:g} \mid \theta,\beta)\, f_g(\theta) d(\theta) \\ &\approx \prod_g \prod_{j:g} \Sigma_k\, P(x_{j:g} \mid \theta = X_k, \beta)\, A_g(X_k). \end{aligned} \tag{1}$$

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 $\mathrm{x}_{j:g}$ 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., $A_g(X_k) = f_g(X_k) / \sum_k f_g(X_k)$). In the ECLS-K:2011, each round of data collection within-grade (e.g., fall and spring second grade) 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 $\hat{N}_k = \sum_g \hat{N}_{gk}$). 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 $\hat{r}_{ik} = \sum_g \hat{r}_{gik}$). 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

[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.

correct. These counts and percentages were checked, item by item, against the statistics generated from the scored, reordered data file to confirm that the correct input file was used and that the information it contained was read correctly by the PARSCALE program.

Another step taken for quality assurance, in addition to verifying the accuracy of the data and computations, was to evaluate the extent to which the scoring model appropriately represented the information in the whole item pool. The *r*-biserials produced in the classical item analysis steps showed the relationship of each test item with the rest of the form on which it appeared. Similarly, the IRT *a* parameter 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, FIRST-GRADE, AND SECOND-GRADE READING, MATHEMATICS, AND SCIENCE IRT ITEM PARAMETERS

Table C-1. ECLS-K:2011 kindergarten, first-grade, and second-grade reading IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, and 2012–13

	Test form(s)										
Item	Kindergarten	1st grade	2nd grade	a[1]	b[2]	c[3]	d1[4]	d2[4]	d3[4]	d4[4]	d5[4]
Phonemic Awareness #1	EBRS	EBRS	†	0.5371	-1.8562	0.0606	†	†	†	†	†
Phonemic Awareness #2	EBRS	EBRS	†	0.5234	-1.0635	0.0154	†	†	†	†	†
Beginning Sounds Item #1	EBRS	EBRS	†	0.7689	-0.0091	0	†	†	†	†	†
Sight Words #1	EBRS	EBRS	†	1.7906	0.1031	0	†	†	†	†	†
Phonemic Awareness #3	R2	R2	†	1.0183	0.2797	0	†	†	†	†	†
Phonemic Awareness #4	R2	R2	†	1.5949	0.2489	0	†	†	†	†	†
Phonemic Awareness #5	R2	R2	†	0.9303	0.4845	0	†	†	†	†	†
Phonemic Awareness #6	R2	R2	†	0.7692	1.1965	0	†	†	†	†	†
Segmentation #1	R2	R2	†	1.2539	-0.5008	0	†	†	†	†	†
Segmentation #2	R2	R2	†	0.6034	1.2761	0	†	†	†	†	†
Sight Words #2	R2	R2	R	1.9581	0.4797	0	†	†	†	†	†
Sight Words #3	R2	R2	R	1.84	0.5463	0	†	†	†	†	†
Sight Words #4	R2	R2	R	1.9057	0.8161	0	†	†	†	†	†
Sight Words #5	R2	R2	R	1.728	0.8724	0	†	†	†	†	†
Locate/Recall #1	R2	R2	R	2.5774	0.8669	0.1785	†	†	†	†	†
Locate/Recall #2	R2	R2	R	1.4535	0.8062	0.2328	†	†	†	†	†
Locate/Recall #3	R2	R2	R	2.5883	0.9136	0.1875	†	†	†	†	†
Locate/Recall #4	R2	R2	R	2.0236	1.2187	0.2227	†	†	†	†	†
Print Convention #1	L	†	†	1.2052	-2.4953	0	†	†	†	†	†
Letter Recognition Item #1	L	†	†	1.0636	-1.7561	0.13	†	†	†	†	†
Print Convention #2	L, M	L	†	1.2133	-1.8093	0	†	†	†	†	†
Ending Sounds Item #1	L, M, H	L	L	0.8522	-0.4708	0	†	†	†	†	†
Vocabulary #1	L	†	†	0.9051	-2.5931	0.2323	†	†	†	†	†
Vocabulary #2	L	†	†	1.152	-1.7705	0.1376	†	†	†	†	†
Vocabulary #3	L	†	†	0.9149	-1.2083	0.1492	†	†	†	†	†
Vocabulary #4	L, M	L	†	0.7277	-1.1787	0.0516	†	†	†	†	†
Vocabulary #5	L	†	†	1.0775	-0.8517	0.2038	†	†	†	†	†
Vocabulary #6	M	L	†	0.4266	-0.5308	0.1289	†	†	†	†	†
Vocabulary #7	M	L	†	0.453	0.7778	0.2001	†	†	†	†	†
Vocabulary #8	M	L	†	0.4171	1.1494	0.2238	†	†	†	†	†
Sight Words #6	M	L	†	1.5703	-0.0937	0	†	†	†	†	†
Sight Words #7	M	L	M	2.1302	0.4074	0	†	†	†	†	†
Sight Words #8	M	L	†	1.7403	0.9928	0	†	†	†	†	†
Locate/Recall #5	H	M	L, M	2.9701	1.3067	0.2288	†	†	†	†	†
Locate/Recall #6	H	M	L, M	4.7023	1.4283	0.2018	†	†	†	†	†
Locate/Recall #7	H	M	L, M	4.0502	1.5296	0.2377	†	†	†	†	†
Locate/Recall #8	H	M	L, M	3.5604	1.5939	0.1718	†	†	†	†	†
Sight Words #9	H	M	L	3.9597	1.1431	0	†	†	†	†	†
Sight Words #10	H	M	L, M	4.64	1.1888	0	†	†	†	†	†
Sight Words #11	H	M	L, M	2.9831	1.5463	0	†	†	†	†	†

See notes at end of table.

Table C-1. ECLS-K:2011 kindergarten, first-grade, and second-grade reading IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, and 2012–13—Continued

	Test form(s)										
Item	Kindergarten	1st grade	2nd grade	a[1]	b[2]	c[3]	d1[4]	d2[4]	d3[4]	d4[4]	d5[4]
Sight Words #12	H	M	L	2.1881	1.9746	0	†	†	†	†	†
Letter Recognition Testlet (5 items)	EBRS	EBRS	†	1.0036	-1.7366	0	0.6579	-0.126	-0.2207	-0.3277	0.0165
Letter Sounds Testlet (2 items)	EBRS	EBRS	†	1.3854	-1.6977	0	0.223	-0.223	†	†	†
Beginning Sounds Testlet (4 items)	EBRS	EBRS	†	0.5339	-0.8943	0	0.398	-0.2137	0.1472	-0.3315	†
Ending Sounds Testlet (3 items)	EBRS	EBRS	†	0.6483	-0.1052	0	0.4848	-0.1037	-0.3811	†	†
Blending Testlet (2 items)	R2	R2	†	1.005	0.0575	0	-0.0015	0.0015	†	†	†
Phonemic Substitution Testlet (2 items)	R2	R2	†	0.8286	0.4072	0	0.117	-0.117	†	†	†
Rhyming Testlet (2 items)	R2	R2	†	0.5144	0.7959	0	-0.4573	0.4573	†	†	†
Passage 1	L	†	†	0.7278	-1.7143	0	0.7222	-0.4407	-0.2815	†	†
Letter Sounds Testlet (2 items)	L, M	L	†	0.3985	-1.7034	0	-1.3124	1.3124	†	†	†
Passage 2	L, M	L	†	0.5915	-0.411	0	0.5727	0.0021	-0.5748	†	†
Passage 3	H	M	L	1.766	1.1132	0	0.2618	0.1796	-0.0885	-0.3529	†
Passage 4	H	M	L, M	1.4967	1.9148	0	0.2916	0.1116	-0.0067	-0.3965	†
Passage 5	H	M	L	1.6211	2.1653	0	0.3614	0.2675	-0.1874	-0.4415	†
Sight Words #13	†	R2	†	1.3705	2.1092	0	†	†	†	†	†
Sight Words #14	†	R2	†	1.9687	2.3225	0	†	†	†	†	†
Sight Words #15	†	R2	R	1.4282	3.2471	0	†	†	†	†	†
Interpret/Integrate #1	†	R2	R	2.1184	1.117	0.4058	†	†	†	†	†
Locate/Recall #9	†	R2	R	1.5466	0.4787	0.0191	†	†	†	†	†
Locate/Recall #10	†	R2	R	1.3322	1.5486	0.1157	†	†	†	†	†
Vocabulary #9	†	H	M, H	2.7649	2.8161	0.2624	†	†	†	†	†
Vocabulary #10	†	H	M, H	1.8138	2.6225	0.3121	†	†	†	†	†
Vocabulary #11	†	H	M, H	3.9176	2.9737	0.3179	†	†	†	†	†
Passage 6	†	R2	R	1.0168	1.5075	0	0.884	0.1566	-0.3153	-0.7252	†
Passage 7	†	H	M, H	0.9601	1.924	0	0.7622	0.2494	-0.2343	-0.7773	†
Passage 8	†	H	M, H	0.6813	2.4827	0	1.2305	0.5716	-0.0503	-0.681	-1.0709
Passage 9	†	H	H	0.3605	3.3193	0	0.9919	-0.0228	0.1022	-1.0712	†
Blending Item #1	†	†	R	0.754	-0.4445	0	†	†	†	†	†
Segmentation Item #1	†	†	R	1.1772	0.4499	0	†	†	†	†	†
Locate/Recall #11	†	†	R	1.6567	1.2278	0.1002	†	†	†	†	†
Locate/Recall #12	†	†	R	1.2357	1.5655	0.0311	†	†	†	†	†
Locate/Recall #13	†	†	R	2.302	1.9944	0.1807	†	†	†	†	†
Sight Words #16	†	†	R	2.2787	2.3826	0	†	†	†	†	†
Sight Words #17	†	†	R	1.6609	3.2184	0	†	†	†	†	†
Sight Words #18	†	†	R	2.0063	3.1613	0	†	†	†	†	†
Passage 10	†	†	R	0.8988	2	0	0.9688	0.4437	-0.0089	-0.4849	-0.9186

[1] Item Response Theory (IRT) discrimination parameter.
[2] Item Response Theory (IRT) difficulty parameter.
[3] Item Response Theory (IRT) guessing parameter.
[4] Item Response Theory (IRT) threshold parameter.

NOTE: † = not administered, EBRS = Early Basic Reading Skills form, R2 = Router 2, R = routing form, L = low second-stage form, M = middle second-stage form, H = high second-stage form.

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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

Table C-2. ECLS-K:2011 kindergarten, first-grade, and second-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13

	Test form(s)					
Item	Kinder-garten	1st grade	2nd grade	a[1]	b[2]	c[3]
Geometry #1	R	†	†	0.6659	-2.36571	0.6
Number Sense - Number Recognition #1	R	†	†	1.51363	-2.27465	0.00718
Number Sense - Number Recognition #2	R	R	†	1.29456	-1.17058	0.00718
Number Sense - Number Recognition #3	R	R	R	1.41898	-0.1223	0.00718
Measurement #1	R	R	†	0.82477	-1.03543	0.28113
Patterns #1	R	R	†	0.96218	0.74949	0.26163
Number Sense - Counting #1	R	R	R	0.94957	-0.89364	0.00713
Data Analysis #1	R	R	†	0.58826	-1.63275	0.00717
Number Sense - Ordinality	R	R	R	1.27	-0.16529	0.0073
Number Sense - Sequencing #1	R	R	R	1.36852	-0.43561	0.00726
Patterns #2	R	R	R	1.26978	1.00727	0.00717
Number Sense - Addition/Subtraction #1	R	R	R	0.98742	0.30492	0.00729
Number Sense - Addition/Subtraction #2	R	R	R	0.90777	0.03027	0.00732
Number Sense - Addition/Subtraction #3	R	†	†	0.99228	0.31862	0.00766
Number Sense - Addition/Subtraction #4	R	R	R	1.09743	0.38372	0.00725
Number Sense - Addition/Subtraction #5	R	R	R	1.24458	0.67093	0.17163
Number Sense - Addition/Subtraction #6	R	R	R	1.45572	0.89732	0.00751
Number Sense - Addition/Subtraction #7	R	†	†	1.02309	0.98404	0.00709
Number Sense - Counting #2	L	†	†	0.45374	-4.69026	0.00717
Number Sense - Counting #3	L	†	†	0.8128	-2.66617	0.00717
Number Sense - Counting #4	L	†	†	0.46608	-2.44108	0.25882
Number Sense - Counting #5	L	†	†	0.73313	-0.87309	0.00717
Number Sense - Counting #6	L	†	†	1.39015	-2.17168	0.00718
Number Sense - Counting #7	L	†	†	1.21022	-0.84371	0.15579
Number Sense - Fewer/More #1	L	†	†	0.47977	-1.04764	0.46644
Number Sense - Fewer/More #2	L	†	†	0.32665	-0.66772	0.00718
Number Sense - Fewer/More #3	L	†	†	0.86728	-1.14506	0.00718
Number Sense - Fewer/More #4	L	†	†	0.47875	0.09545	0.00718
Data Analysis #2	L, M	L	†	0.80133	-1.55796	0.00716
Data Analysis #3	L, M	L	†	0.80706	-1.92272	0.00716
Data Analysis #4	L	†	†	0.79346	-2.42697	0.00718
Number Sense - Addition/Subtraction #8	L, M	L	†	0.7149	-1.22448	0.0072
Number Sense - Number Recognition #4	L	†	†	1.32963	-1.78673	0.00718
Number Sense - Number Recognition #5	L, M, H	L, M	L	1.32987	-0.47767	0.00714

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, and second-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13—Continued

	Test form(s)					
Item	Kinder-garten	1st grade	2nd grade	a[1]	b[2]	c[3]
Measurement #2	L, M	L	†	0.88961	-1.12755	0.27931
Number Sense - Number Recognition #6	M	L	†	0.99815	-0.36646	0.20174
Number Sense - Fewer/More #5	M	L	†	0.64178	-0.19626	0.21673
Number Sense - Fewer/More #6	M	L	†	0.6781	0.21172	0.1225
Patterns #3	M, H	L, M	L	1.03496	0.54848	0.29342
Geometry #2	M	L	†	0.65623	-0.13244	0.17536
Patterns #4	M	L	†	0.7854	-0.21601	0.27445
Number Sense - Counting #8	M	L	†	0.60259	-0.65709	0.13093
Number Sense - Addition/Subtraction #9	M	L	†	1.04225	0.38229	0.00717
Number Sense - Addition/Subtraction #10	M	L	†	1.07366	0.0983	0.12034
Number Sense - Addition/Subtraction #11	M	L	†	0.68746	-0.1822	0.08315
Number Sense - Estimation #1	M	L	†	0.87449	0.12913	0.15758
Number Sense - Addition/Subtraction #12	M, H	L, M	L	2.6416	0.13706	0.00736
Number Sense - Addition/Subtraction #13	M	L	†	3.73736	0.21493	0.00728
Number Sense - Addition/Subtraction #14	M, H	L, M	L	2.06009	0.44219	0.00703
Number Sense - Addition/Subtraction #15	M, H	L, M	L	1.08845	1.61394	0.0157
Number Sense - Addition/Subtraction #16	M, H	L, M	L	2.44748	1.65596	0.00669
Number Sense - Counting #9	M	L	†	1.46858	0.82267	0.00712
Number Sense - Sequencing #2	M	L	†	1.5455	0.72402	0.00717
Number Sense - Addition/Subtraction #17	H	M	L	0.74424	-0.43912	0.00717
Number Sense - Addition/Subtraction #18	H	M	L	1.89653	0.87868	0.00716
Number Sense - Addition/Subtraction #19	H	M	L	1.76243	1.12956	0.00715
Patterns #5	H	M	L	0.69102	-0.10694	0.19717
Patterns #6	H	M	L	1.74515	0.83431	0.00717
Patterns #7	H	M	L	2.10658	0.92751	0.00717
Patterns #8	H	M, H	L, M	1.1967	1.34231	0.00716
Data Analysis #5	H	M	L	1.596	1.57863	0.0072
Number Sense - Addition/Subtraction #20	H	M	L	1.28066	0.44528	0.00717
Number Sense - Addition/Subtraction #21	H	M	L	1.4336	1.72465	0.00717
Number Sense - Addition/Subtraction #22	H	M	L	1.91524	1.74505	0.00712
Number Sense - Addition/Subtraction #23	H	M	L	1.90998	1.76027	0.00713
Number Sense - Multiplication/Division #1	H	M, H	L, M	1.6469	1.48735	0.0072
Number Sense - Multiplication/Division #2	H	M, H	L, M	2.26105	1.68549	0.00725
Number Sense - Multiplication/Division #3	H	M, H	L, M	2.00911	1.75329	0.00726

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, and second-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13—Continued

	Test form(s)					
Item	Kinder-garten	1st grade	2nd grade	a[1]	b[2]	c[3]
Number Sense - Estimation #2	H	M, H	L, M	1.5558	2.06217	0.00723
Number Sense - Addition/Subtraction #24	H	M, H	L, M	1.81985	1.832	0.00721
Data Analysis #6	H	M, H	L, M	0.80023	2.35816	0.00718
Number Sense - Addition/Subtraction #25	H	M, H	L, M	2.66354	1.96964	0.00711
Number Sense - Multiplication/Division #4	H	M, H	L, M	1.09912	2.00999	0.00718
Number Sense - Money #1	H	M, H	L, M	1.42507	2.39731	0.00728
Number Sense - Money #2	H	M, H	L, M	2.81439	2.19963	0.00708
Number Sense - Multiplication/Division #5	†	R	†	1.43708	2.09036	0.00009
Number Sense - Money #3	†	R	R	1.85251	3.48864	0.00018
Number Sense - Addition/Subtraction #26	†	R	R	1.34077	4.47342	0.0001
Number Sense - Addition/Subtraction #27	†	H	M, H	1.95781	2.69652	0.00008
Patterns #9	†	H	M, H	1.69861	2.8698	0.00008
Patterns #10	†	H	M, H	1.08954	2.67794	0.00008
Number Sense - Addition/Subtraction #28	†	H	M, H	2.20744	3.00318	0.00008
Number Sense - Multiplication/Division #6	†	H	M	0.9157	2.03957	0.06853
Measurement #3	†	H	M, H	0.75221	2.93155	0.00008
Number Sense - Addition/Subtraction #29	†	H	M	0.84013	2.72308	0.00008
Geometry #3	†	H	M	0.73334	2.64631	0.0343
Number Sense - Multiplication/Division #7	†	H	M, H	1.18803	3.18923	0.26031
Number Sense - Fractions #1	†	H	M, H	1.24835	4.14001	0.13195
Measurement #4	†	H	M, H	1.70707	3.521	0.00008
Number Sense - Multiplication/Division #8	†	H	M	1.09754	3.10893	0.00008
Number Sense - Money #4	†	H	M, H	1.18062	3.7016	0.00008
Data Analysis #7	†	H	M, H	1.02484	3.94844	0.21966
Number Sense - Multiplication/Division #9	†	H	M	0.93676	4.02794	0.00008
Number Sense - Fractions #2	†	H	M	1.78864	4.20016	0.0631
Number Sense - Multiplication/Division #10	†	H	M	1.21479	3.58929	0.15005
Number Sense - Multiplication/Division #11	†	H	M, H	0.99876	3.76087	0.00008
Data Analysis #8	†	†	R	0.96496	1.11683	0.35782
Number Sense - Addition/Subtraction #30	†	†	R	0.85057	1.44874	0.00001
Geometry #4	†	†	R	0.97929	2.18583	0.21441
Data Analysis #9	†	†	R	0.98805	1.98932	0.00001
Number Sense - Estimation #3	†	†	R	0.56988	1.95057	0.10818
Data Analysis #10	†	†	R	1.27215	3.08893	0.00001

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, and second-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13—Continued

Item	Test form(s)					
	Kinder-garten	1st grade	2nd grade	a[1]	b[2]	c[3]
Number Sense - Fractions #3	†	†	R	0.51507	3.73762	0.00001
Number Sense - Multiplication/Division #12	†	†	R	1.49389	4.18346	0.00003
Number Sense - Addition/Subtraction #31	†	†	H	1.32332	2.42271	0.00001
Number Sense - Comparison #1	†	†	H	1.41113	2.61065	0.2454
Number Sense - Addition/Subtraction #32	†	†	H	0.9434	2.74463	0.00001
Number Sense - Money #5	†	†	H	1.57844	3.21939	0.00001
Measurement #5	†	†	H	1.72306	3.65849	0.00001
Measurement #6	†	†	H	1.95838	3.6152	0.00001
Number Sense - Money #6	†	†	H	1.84105	3.68433	0.00001
Number Sense - Reasoning	†	†	H	1.50118	4.27577	0.00001
Number Sense - Comparison #2	†	†	H	0.94285	4.60037	0.00001

[1] Item Response Theory (IRT) discrimination parameter.
[2] Item Response Theory (IRT) difficulty parameter.
[3] Item Response Theory (IRT) guessing parameter.
NOTE: † = not administered, R = routing form, L = low second-stage form, M = middle second-stage form, H = high second-stage form.
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, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013.

Table C-3. ECLS-K:2011 kindergarten, first-grade, and second-grade science IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, and 2012–13

	Test form(s)					
Item	Kindergarten	1st grade	2nd grade	a[1]	b[2]	c[3]
Life Science #1	R	†	†	1.22028	-1.12805	0.825
Physical Science #1	R	†	†	0.59425	-0.25402	0.56292
Earth Science #1	R	†	†	0.55531	-1.56503	0.40063
Scientific Inquiry #1	R	L, M	L	0.73863	-1.16655	0.00001
Life Science #2	R	R	R	1.17529	-1.09259	0.00001
Earth Science #2	R	†	†	0.93555	-0.69703	0.00001
Earth Science #3	R	R	R	0.78569	-0.84103	0.00001
Physical Science #2	R	L, M, H	L, M	1.06078	-0.09729	0.16824
Physical Science #3	R	R	R	0.83743	-0.28915	0.19376
Scientific Inquiry #2	R	L	†	0.79794	-0.88006	0.00001
Life Science #3	R	R	R	1.08132	0.40997	0.00001
Scientific Inquiry #3	R	L	†	0.85499	0.43745	0.30835
Life Science #4	R	R	R	0.76793	0.81077	0.39703
Scientific Inquiry #4	R	R	R	0.91231	0.5512	0.14747
Scientific Inquiry #5	R	R	R	0.87386	3.06918	0.21654
Physical Science #4	R	L	†	1.36161	1.27296	0.16679
Physical Science #5	R	R	R	1.07314	1.86303	0.05265
Earth Science #4	R	R	†	1.14704	1.60448	0.13601
Earth Science #5	R	M	L	1.17355	1.59889	0.2286
Life Science #5	R	R	†	0.84546	1.44217	0.28632
Earth Science #6	†	R	R	0.55347	-1.44121	0.10907
Life Science #6	†	R	†	0.99675	0.64508	0.00001
Life Science #7	†	R	†	0.87301	0.30981	0.00001
Earth Science #7	†	R	†	1.06675	2.22400	0.08302
Earth Science #8	†	R	R	0.01021	70.59083	0.00086
Earth Science #9	†	L	†	0.89748	-2.30713	0.00001
Physical Science #6	†	L, M	L	0.82732	-1.87128	0.00001
Physical Science #7	†	L, M	L	0.91025	-0.87593	0.05771
Physical Science #8	†	L, M	L	0.64199	-1.03951	0.08177
Life Science #8	†	L, M, H	L, M	0.69868	-0.46413	0.00001
Life Science #9	†	L, M, H	L, M	0.89280	-0.58470	0.26815
Earth Science #10	†	L, M	L	1.04379	0.66691	0.09713
Life Science #10	†	L, M, H	L, M	1.03933	0.60494	0.00001
Scientific Inquiry #6	†	L	†	0.72614	2.65844	0.27612
Scientific Inquiry #7	†	M	L	0.66814	-0.56039	0.15313
Life Science #11	†	M, H	L, M	0.62425	-0.30107	0.10858
Physical Science #9	†	M, H	L, M, H	0.91177	1.15807	0.13653
Earth Science #11	†	M, H	L, M, H	1.00312	1.49124	0.44506
Life Science #12	†	M	L	1.31838	1.36822	0.12901
Scientific Inquiry #8	†	H	M	0.34545	-0.46448	0.18829
Life Science #13	†	H	M, H	0.87400	1.26139	0.15342
Scientific Inquiry #9	†	H	M, H	0.75022	1.88149	0.26162

See notes at end of table.

Table C-3. ECLS-K:2011 kindergarten, first-grade, and second-grade science IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, and 2012–13—Continued

	Test form(s)					
Item	Kindergarten	1st grade	2nd grade	a[1]	b[2]	c[3]
Earth Science #12	†	H	M, H	0.76484	1.59826	0.00001
Earth Science #13	†	H	M	0.82958	2.21811	0.03335
Scientific Inquiry #10	†	H	M, H	1.83441	2.75588	0.14182
Life Science #14	†	H	M, H	0.71855	2.05893	0.19323
Life Science #15	†	H	M, H	0.85422	3.03732	0.00001
Earth Science #14	†	†	R	0.57488	-3.32234	0.00001
Physical Science #10	†	†	R	0.95102	-0.34155	0.00001
Physical Science #11	†	†	R	0.42000	0.44635	0.00001
Scientific Inquiry #11	†	†	R	1.60607	1.35998	0.33369
Scientific Inquiry #12	†	†	R	0.92008	0.64714	0.29409
Scientific Inquiry #13	†	†	R	0.98965	1.60451	0.00001
Physical Science #12	†	†	R	1.73670	1.40134	0.14390
Earth Science #15	†	†	R	0.94518	2.78342	0.00002
Earth Science #16	†	†	R	1.06631	2.59025	0.15499
Scientific Inquiry #14	†	†	H	0.91895	0.86546	0.21987
Scientific Inquiry #15	†	†	H	0.97383	1.15051	0.24134
Physical Science #13	†	†	H	0.73425	1.98890	0.25475
Physical Science #14	†	†	H	0.65112	1.43852	0.00001
Scientific Inquiry #16	†	†	H	1.25631	2.37446	0.29912
Life Science #16	†	†	H	1.45139	2.52856	0.13140
Earth Science #17	†	†	H	1.84555	2.72302	0.15018
Physical Science #15	†	†	H	1.08580	2.58192	0.00001

[1] Item Response Theory (IRT) discrimination parameter.

[2] Item Response Theory (IRT) difficulty parameter.

[3] Item Response Theory (IRT) guessing parameter.

NOTE: † = not administered, R = routing form, L = low second-stage form, M = middle second-stage form, H = high second-stage form.

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, fall 2011, spring 2012, fall 2012, and spring 2013.

Table C-4. ECLS-K kindergarten and first-grade Spanish early reading skills IRT item parameters on the concurrent kindergarten and first-grade scale: School years 2010–11 and 2011–12

Item	a[1]	b[2]	c[3]
Print Convention #1	0.88502	-2.08767	0.00001
Vocabulary #1	0.78433	-2.24613	0.00001
Vocabulary #2	1.02292	-1.57189	0.00001
Print Convention #2	1.0474	-0.92211	0.00001
Print Convention #3	0.88406	-1.61907	0.00001
Print Convention #4	0.94611	-0.6039	0.00001
Print Convention #5	0.81669	0.10487	0.00001
Print Convention #6	0.82854	0.64742	0.00001
Print Convention #7	0.93081	0.08579	0.00001
Letter Recognition #1	0.39505	-0.33253	0.24178
Letter Recognition #2	1.94235	-0.09753	0.00001
Letter Recognition #3	2.08681	-0.30194	0.00001
Letter Recognition #4	2.30784	-0.13945	0.00001
Letter Recognition #5	2.07263	-0.34522	0.00001
Letter Sounds #1	1.27061	-0.92847	0.00001
Letter Sounds #2	0.84263	-0.00613	0.22845
Letter Sounds #3	0.87871	-0.04222	0.20066
Beginning Sounds #1	1.00011	-0.07731	0.00001
Beginning Sounds #2	0.85477	0.01277	0.00001
Vocabulary #3	0.47445	-0.07183	0.29754
Vocabulary #4	0.59058	-0.79525	0.24567
Vocabulary #5	0.54751	1.27786	0.24999
Vocabulary #6	1.0571	0.67653	0.19695
Vocabulary #7	0.84802	1.33362	0.29314
Sight Words #1	2.62889	0.52799	0.00001
Sight Words #2	1.98593	0.49075	0.00001
Sight Words #3	5.73248	0.7295	0.00001
Sight Words #4	4.01508	1.04087	0.00001
Sight Words #5	6.53303	0.79238	0.00001
Sight Words #6	5.40942	0.94802	0.00001
Sight Words #7	5.89665	0.99508	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, spring 2011, fall 2011, and spring 2012.

This page intentionally left blank.

APPENDIX D

PLOTS OF IRT THETA AND SCALE SCORE DISTRIBUTIONS IN READING, MATHEMATICS, AND SCIENCE

Figure D-1. IRT theta score distributions in reading for school years 2010–11, 2011–12, and 2012–13

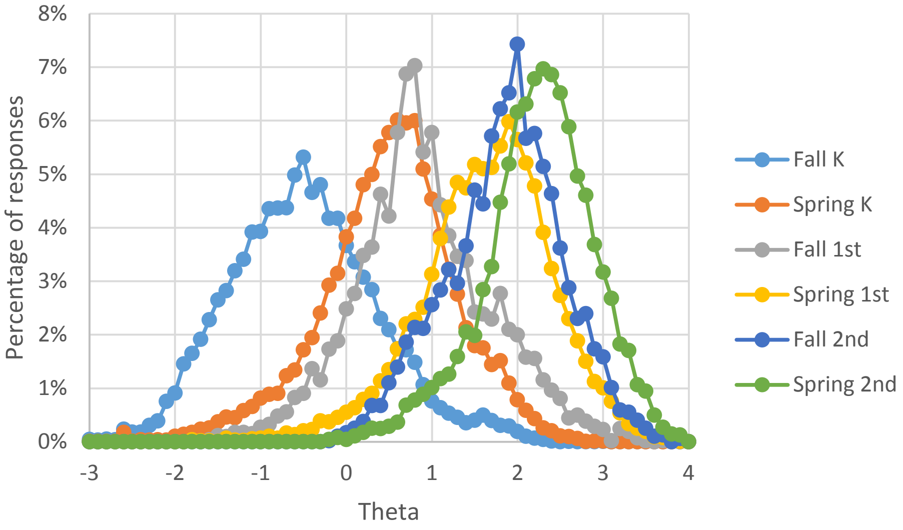


Figure D-2. IRT scale score distributions in reading for school years 2010–11, 2011–12, and 2012–13

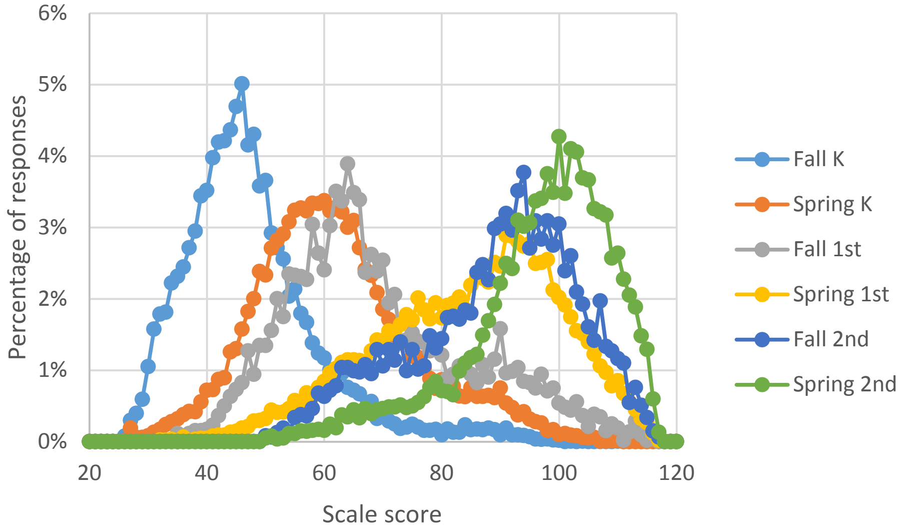


Figure D-3. IRT theta score distributions in mathematics for school years 2010–11, 2011–12, and 2012-13

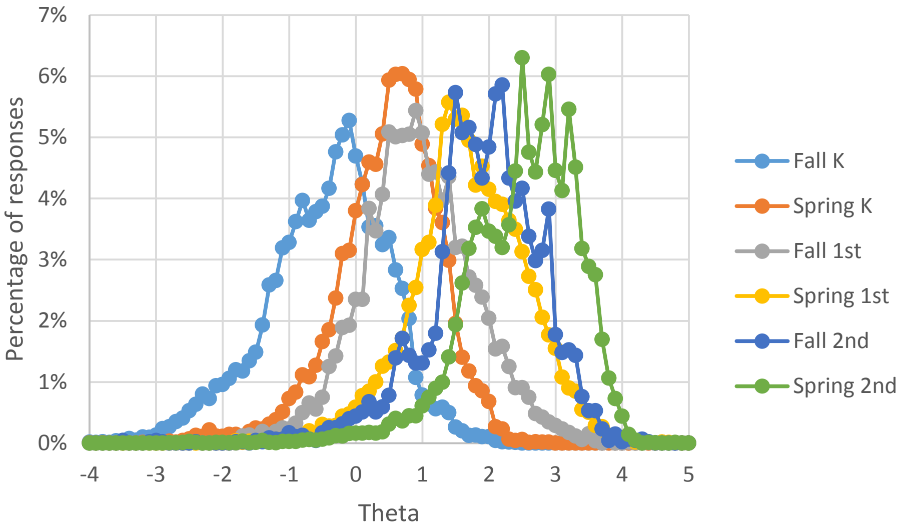


Figure D-4. IRT scale score distributions in mathematics for school years 2010–11, 2011–12, and 2012–13

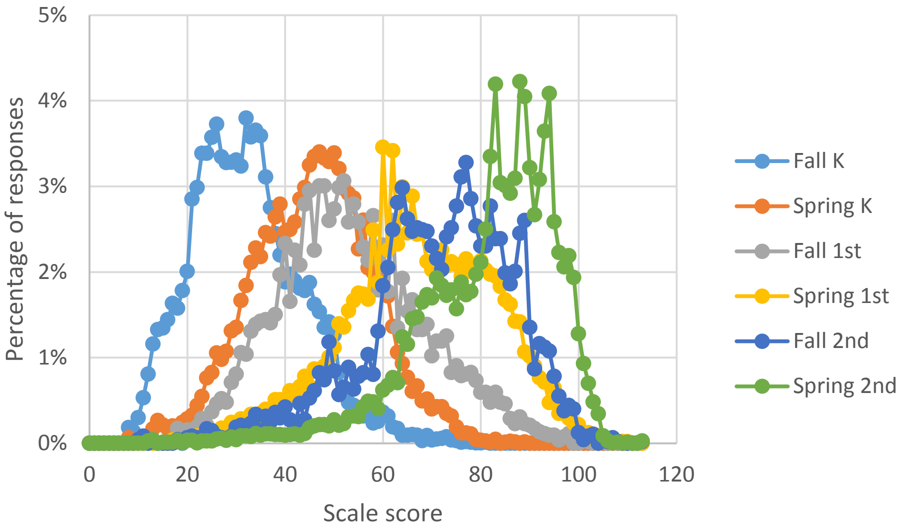


Figure D-5. IRT theta score distributions in science for school years 2010–11, 2011–12, and 2012–13

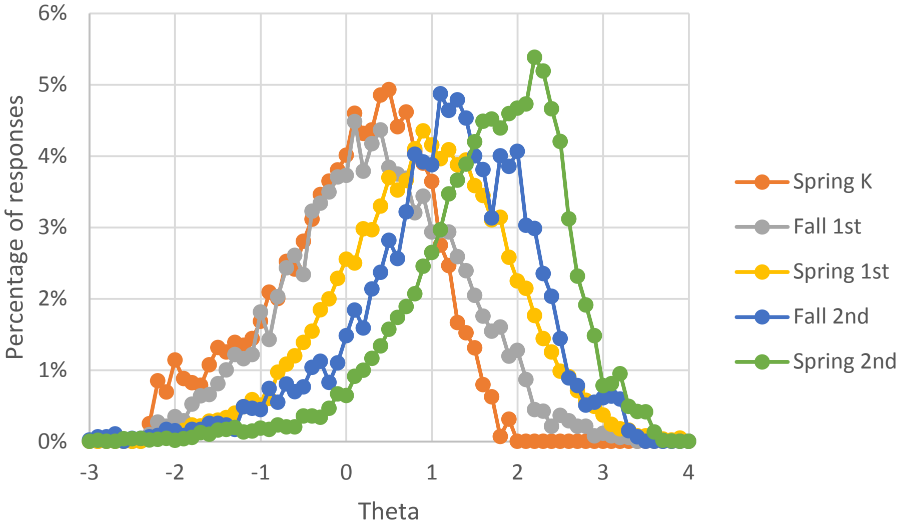
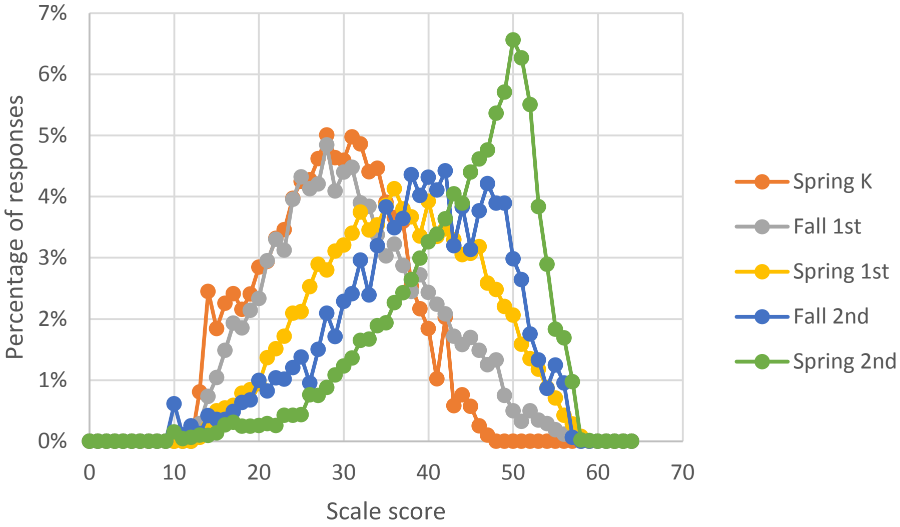


Figure D-6. IRT scale score distributions in science for school years 2010–11, 2011–12, and 2012–13



This page intentionally left blank.