IES Institute of Education Sciences

Early Childhood Longitudinal Study, Kindergarten Class of 2010-11

Third-Grade, Fourth-Grade, and Fifth-Grade Psychometric Report

Technical/Methodological Report

NCES 2020-123
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Statistics at IES

Early Childhood Longitudinal Study, Kindergarten Class of 2010–11

Third-Grade, Fourth-Grade, and Fifth-Grade Psychometric Report

MARCH 2020

Michelle Najarian
Educational Testing Service (ETS)

Karen Tourangeau
Christine Nord
Kathleen Wallner-Allen
Nancy Vaden-Kiernan
Westat

Gail M. Mulligan
Project Officer
National Center for Education Statistics

NCES 2020-123
U.S. Department of Education

IES 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
550 12th Street SW
Washington, DC 20202

March 2020

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-14-C-0119 with Westat and the Educational Testing Service. 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., Wallner-Allen, K., and Vaden-Kiernan, N. (2019). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), Third-Grade, Fourth-Grade, and Fifth-Grade Psychometric Report* (NCES 2020-123). U.S. Department of Education. Institute of Education Sciences. Washington, DC: National Center for Education Statistics. Retrieved [Month day, year] from https://nces.ed.gov/pubsearch.

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

CONTENTS

Chapter				Page
LIST OF APPENDIXES				ix
LIST OF TABLES				ix
LIST OF FIGURES				xix
LIST OF EXHIBITS				xx
1	INTRODUCTION			1-1
	1.1	The Third-Grade, Fourth-Grade, and Fifth-Grade Data Collections		1-3
	1.2	Data Collection Instruments and Administration of Assessments		1-4
	1.3	Contents of Report		1-7
2	OVERVIEW OF THE CONTENT COVERAGE AND ADMINISTRATION OF THE DIRECT COGNITIVE ASSESSMENT INSTRUMENTS			2-1
	2.1	Content Coverage of the Cognitive Assessments		2-2
		2.1.1	Reading Test Specifications	2-2
			2.1.1.1 Basic Reading Skills	2-3
			2.1.1.2 Vocabulary	2-4
			2.1.1.3 Comprehension	2-5
			2.1.1.4 Continuity Between the ECLS-K and the ECLS-K:2011 Reading Frameworks	2-6
		2.1.2	Mathematics Test Specifications	2-6
			2.1.2.1 Number Properties and Operations	2-7
			2.1.2.2 Measurement	2-7
			2.1.2.3 Geometry	2-8
			2.1.2.4 Data Analysis and Probability	2-8
			2.1.2.5 Algebra	2-8
		2.1.3	Science Test Specifications	2-9
			2.1.3.1 Scientific Inquiry	2-9
			2.1.3.2 Physical Science	2-9
			2.1.3.3 Life Science	2-10
			2.1.3.4 Earth and Space Science	2-10
	2.2	Item and Time Allocation Across and Within Subject Areas		2-10

Chapter				Page
	2.3	Mode of Administration		2-13
	2.4	Inclusion of Children With Diverse Language Backgrounds and Language of Administration		2-15
3	ANALYSIS METHODOLOGY			3-1
	3.1	Quality Control Procedures		3-1
	3.2	Overview of Item Response Theory (IRT)		3-4
		3.2.1	Dichotomous Item Calibration	3-6
		3.2.2	Polytomous Item Calibration	3-10
		3.2.3	Calculation of IRT Scale Scores	3-13
		3.2.4	Selection of an IRT Model	3-13
		3.2.5	Evaluating Items Using Empirical Item Characteristic Curves	3-15
		3.2.6	Item Information and Measurement Precision	3-18
		3.2.7	Item Response Theory Estimation Using PARSCALE	3-22
	3.3	Construct Validity: Assessing Dimensionality		3-24
	3.4	Group Differences in Item Functioning		3-25
	3.5	Development of the Kindergarten Through Fifth-Grade Longitudinal Scale		3-29
		3.5.1	Evaluating Common Item Functioning	3-30
		3.5.2	Concurrent Calibration and Chain Linking	3-31
			3.5.2.1 Methodology Approaches	3-32
			3.5.2.2 Comparison of Approaches and Methodology Selection	3-34
			3.5.2.3 Chain-Linking Methodology	3-35
		3.5.3	Computation of Final Scale Scores	3-36
4	DEVELOPMENT OF THE TWO-STAGE COGNITIVE ASSESSMENT TEST FORMS			4-1
	4.1	Development of the Item Pool		4-1
		4.1.1	Field-test Design	4-1
		4.1.2	Methods Used to Analyze Data for Design of the National Assessments	4-2
		4.1.3	Criteria Guiding the Selection of Items for the National Assessments	4-5

Chapter Page

4.2 Reading 4-8

4.2.1 Estimated Ability Levels for the ECLS-K:2011 National Samples and Target Ranges for Item Difficulties 4-8

4.3 Mathematics 4-13

4.3.1 Estimated Ability Levels for the ECLS-K:2011 National Samples and Target Ranges for Item Difficulties 4-13

4.4 Science 4-18

4.4.1 Estimated Ability Levels for the ECLS-K:2011 National Samples and Target Ranges for Item Difficulties 4-18

4.5 Cut Scores Used for Routing 4-22

4.5.1 Reading 4-24
4.5.2 Mathematics 4-25
4.5.3 Science 4-25

4.6 Discontinue Rules 4-26

4.6.1 Reading 4-27
4.6.2 Mathematics 4-27
4.6.3 Science 4-28

5 PSYCHOMETRIC CHARACTERISTICS OF THE ECLS-K:2011 DIRECT COGNITIVE BATTERY 5-1

5.1 Routing of Children Through the National Assessments 5-1

5.1.1 Third-Grade Round 5-2
5.1.2 Fourth-Grade Round 5-3
5.1.3 Fifth-Grade Round 5-4

5.2 Scoring the National Assessment 5-4

5.2.1 Confirmation of IRT Assumptions 5-5
5.2.2 Analysis of Differential Item Functioning (DIF) 5-9

5.2.2.1 Third-Grade Round 5-9

Chapter Page

5.2.2.2 Fourth-Grade Round 5-11
5.2.2.3 Fifth-Grade Round 5-12

5.2.3 Assessment Score Reliability 5-13
5.2.4 Item Response Theory (IRT)-Based Scores Developed for the ECLS-K:2011 5-14

5.2.4.1 Theta and the Standard Error of Measurement (SEM) of Theta 5-15
5.2.4.2 IRT Scale Scores 5-17

5.3 Reading Assessment 5-18

5.3.1 Samples and Associated Statistics for the Third-Grade Round 5-18
5.3.2 Samples and Associated Statistics for the Fourth-Grade Round 5-19
5.3.3 Samples and Associated Statistics for the Fifth-Grade Round 5-20
5.3.4 Score Statistics 5-21
5.3.5 Reliabilities 5-27

5.4 Mathematics Assessment 5-27

5.4.1 Samples and Associated Statistics for the Third-Grade Round 5-27
5.4.2 Samples and Associated Statistics for the Fourth-Grade Round 5-28
5.4.3 Samples and Associated Statistics for the Fifth-Grade Round 5-29
5.4.4 Score Statistics 5-30
5.4.5 Reliabilities 5-36

5.5 Science Assessment 5-36

5.5.1 Samples and Associated Statistics in the Third-Grade Round 5-36
5.5.2 Samples and Associated Statistics in the Fourth-Grade Round 5-37
5.5.3 Samples and Associated Statistics in the Fifth-Grade Round 5-38
5.5.4 Score Statistics 5-39
5.5.5 Reliabilities 5-45

Chapter Page

5.6 Evaluating the Kindergarten Through Fifth-Grade Assessment Battery: Construct Validity and Measurement of Growth 5-46

5.6.1 Is the Difficulty of the Tests Suitable for Children's Ability Levels? 5-46
5.6.2 Do the Tests Measure the Right Content? 5-47
5.6.3 Do the Data Constitute a Cohesive Scale Suitable for Longitudinal Measurement? 5-47

5.6.3.1 Precision of Gain Scores Used to Model Growth 5-49

5.6.4 Relationship of the Cognitive Test Scores to Scores in Different Rounds and Different Subjects 5-50

5.7 Selection and Use of Scores 5-53

5.7.1 Choosing the Appropriate Score for Analysis 5-53
5.7.2 Analytic Considerations for Measuring Gains in the ECLS-K:2011 5-54

6 PSYCHOMETRIC CHARACTERISTICS OF THE EXECUTIVE FUNCTION MEASURES 6-1

6.1 Dimensional Change Card Sort 6-1

6.1.1 Mean Scores for the Dimensional Change Card Sort 6-7
6.1.2 Use of Dimensional Change Card Sort Scores in Longitudinal Analyses 6-10
6.1.3 Dimensional Change Card Sort Data Flags 6-12

6.2 Numbers Reversed 6-15

6.2.1 Mean Scores for Numbers Reversed 6-20
6.2.2 Use of Numbers Reversed Scores in Longitudinal Analyses 6-28
6.2.3 Numbers Reversed Data Flags 6-29

6.3 The NIH Toolbox Flanker Inhibitory Control and Attention Task (Flanker) 6-30

6.3.1 Mean Scores for the Flanker 6-36
6.3.2 Use of Flanker Scores in Longitudinal Analyses 6-38
6.3.3 Flanker Data Flags 6-39

Chapter Page

6.4 Correlations Among Executive Function Measures 6-40
6.5 Examination of Possible Assessor Effects in the Administration of the Executive Function Measures 6-42

6.5.1 Numbers Reversed 6-44
6.5.2 DCCS 6-46
6.5.3 Flanker 6-48
6.5.3 Skew 6-49

7 CHILD QUESTIONNAIRE 7-1

7.1 Self-Description Questionnaire (SDQ) 7-5
7.2 Peer Victimization 7-6
7.3 Social Anxiety/Fear of Negative Evaluation 7-6
7.4 Prosocial Behavior 7-7
7.5 Life Satisfaction 7-7
7.6 Behavioral Engagement 7-8
7.7 Peer Social Support 7-8
7.8 Loneliness 7-9
7.9 Media Usage 7-9
7.10 Pets 7-10
7.11 School Belonging 7-10
7.12 Grit 7-11
7.13 Worry/Stress About School 7-11
7.14 Parental Monitoring 7-12

8 PSYCHOMETRIC CHARACTERISTICS OF THE INDIRECT CHILD MEASURES 8-1

8.1 Teacher Measures 8-2

8.1.1 Teacher-Reported Social Skills 8-3
8.1.2 Teacher-Reported Approaches to Learning Items and Scale 8-14
8.1.3 Teacher-Reported Attentional Focusing and Inhibitory Control: Children's Behavior Questionnaire (CBQ) and Temperament in Middle Childhood Questionnaire (TMCQ) 8-17
8.1.4 Teacher-Reported Student-Teacher Relationship Scale 8-24
8.1.5 Teacher-Reported Working Memory 8-27
8.1.6 Teacher-Reported Measures of the Peer Context and the Child's Relationships with Peers 8-30
8.1.7 Teacher-Reported Perceptions of Child's School Liking 8-34

8.2 Parent Measures 8-35

8.2.1 Parent-Reported Working Memory 8-35

Chapter Page

8.2.2 Parent-Reported Measures of the Child's Friendships and Relationships with Peers 8-38
8.2.3 Parent-Reported School Avoidance 8-40

REFERENCES R-1

List of Appendixes

Appendix

A PLOTS OF ITEM RESPONSE THEORY (IRT) TEST INFORMATION FUNCTIONS FOR READING, MATHEMATICS, AND SCIENCE A-1

B ITEM RESPONSE THEORY (IRT) ESTIMATION USING PARSCALE B-1

C ECLS-K:2011 KINDERGARTEN, FIRST-GRADE, SECOND-GRADE, THIRD-GRADE, FOURTH-GRADE, AND FIFTH-GRADE READING, MATHEMATICS, AND SCIENCE ITEM RESPONSE THEORY (IRT) ITEM PARAMETERS C-1

D PLOTS OF ITEM RESPONSE THEORY (IRT) THETA AND SCALE SCORE DISTRIBUTIONS IN READING, MATHEMATICS, AND SCIENCE D-1

List of Tables

Table

2-1 Reading content categories and target percentages of items, ECLS-K:2011 kindergarten through fifth-grade assessments 2-12

2-2 Mathematics content categories and target percentages of items, ECLS-K:2011 kindergarten through fifth-grade assessments 2-12

2-3 Science content categories and target percentages of items, ECLS-K:2011 kindergarten through fifth-grade assessments 2-13

3-1 Comparison of one-, two-, and three-parameter IRT models of relative model fit (AIC and BIC) 3-14

4-1 Means and standard deviations of estimated reading ability level (theta) for children in third grade 4-9

List of Tables—Continued

Table		Page
4-2	Means and standard deviations of estimated reading ability level (theta) for children in fourth grade	4-9
4-3	Means and standard deviations of estimated reading ability level (theta) for children in fifth grade	4-9
4-4	Peak difficulty ranges for the national third-grade reading assessment, routing plus second stage: ECLS-K:2011	4-10
4-5	Peak difficulty ranges for the national fourth-grade reading assessment, routing plus second stage: ECLS-K:2011	4-10
4-6	Peak difficulty ranges for the national fifth-grade reading assessment, routing plus second stage: ECLS-K:2011	4-11
4-7	Framework targets and items by content area for the national third-grade, fourth-grade, and fifth-grade reading assessments: ECLS-K:2011	4-12
4-8	Dichotomous and polytomous items and score values for the national third-grade, fourth-grade, and fifth-grade reading assessments: ECLS-K:2011	4-13
4-9	Means and standard deviations of estimated mathematics ability level (theta) for children in third grade	4-14
4-10	Means and standard deviations of estimated mathematics ability level (theta) for children in fourth grade	4-15
4-11	Means and standard deviations of estimated mathematics ability level (theta) for children in fifth grade	4-15
4-12	Peak difficulty ranges for the national third-grade mathematics assessment, routing plus second stage: ECLS-K:2011	4-16
4-13	Peak difficulty ranges for the national fourth-grade mathematics assessment, routing plus second stage: ECLS-K:2011	4-16
4-14	Peak difficulty ranges for the national fifth-grade mathematics assessment, routing plus second stage: ECLS-K:2011	4-17
4-15	Framework targets and items by content area for the national third-grade, fourth-grade, and fifth-grade mathematics assessments: ECLS-K:2011	4-18

List of Tables—Continued

Table		Page
4-16	Means and standard deviations of estimated science ability level (theta) for children in third grade	4-19
4-17	Means and standard deviations of estimated science ability level (theta) for children in fourth grade	4-19
4-18	Means and standard deviations of estimated science ability level (theta) for children in fifth grade	4-20
4-19	Peak difficulty ranges for the national third-grade science assessment, routing plus second stage: ECLS-K:2011	4-20
4-20	Peak difficulty ranges for the national fourth-grade science assessment, routing plus second stage: ECLS-K:2011	4-21
4-21	Peak difficulty ranges for the national fifth-grade science assessment, routing plus second stage: ECLS-K:2011	4-21
4-22	Framework targets and items by content area for the national third-grade, fourth-grade, and fifth-grade science assessments: ECLS-K:2011	4-22
5-1	Component analysis percentages by component by domain, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11	5-6
5-2	Component analysis percentages by component by domain, ECLS-K:2011 fall and spring first-grade data collections: School year 2011–12	5-7
5-3	Component analysis percentages by component by domain, ECLS-K:2011 fall and spring second-grade data collections: School year 2012–13	5-7
5-4	Component analysis percentages by component by domain, ECLS-K:2011 spring third-grade data collection: Spring 2014	5-7
5-5	Component analysis percentages by component by domain, ECLS-K:2011 spring fourth-grade data collection: Spring 2015	5-7
5-6	Component analysis percentages by component by domain, ECLS-K:2011 spring fifth-grade data collection: School year 2015–16	5-8

List of Tables—Continued

Table		Page
5-7	Reading assessment differential item functioning, ECLS-K:2011 spring third-grade data collection: Spring 2014	5-10
5-8	Mathematics assessment differential item functioning, ECLS-K:2011 spring third-grade data collection: Spring 2014	5-10
5-9	Science assessment differential item functioning, ECLS-K:2011 spring third-grade data collection: Spring 2014	5-10
5-10	Reading assessment differential item functioning, ECLS-K:2011 spring fourth-grade data collection: Spring 2015	5-11
5-11	Mathematics assessment differential item functioning, ECLS-K:2011 spring fourth-grade data collection: Spring 2015	5-11
5-12	Science assessment differential item functioning, ECLS-K:2011 spring fourth-grade data collection: Spring 2015	5-12
5-13	Reading assessment differential item functioning, ECLS-K:2011 spring fifth-grade data collection: Spring 2016	5-12
5-14	Mathematics assessment differential item functioning, ECLS-K:2011 spring fifth-grade data collection: Spring 2016	5-13
5-15	Science assessment differential item functioning, ECLS-K:2011 spring fifth-grade data collection: Spring 2016	5-13
5-16	Third-grade reading assessment sample, ECLS-K:2011 spring third-grade data collections: Spring 2014	5-19
5-17	Fourth-grade reading assessment sample, ECLS-K:2011 spring fourth-grade data collection: Spring 2015	5-20
5-18	Fifth-grade reading assessment samples, ECLS-K:2011 spring fifth-grade data collection: Spring 2016	5-21
5-19	Reading assessment statistics, by IRT-based score, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	5-22

List of Tables—Continued

Table		Page
5-20	Reading assessment theta statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	5-23
5-21	Reading assessment scale score statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	5-25
5-22	Reading assessment reliabilities for IRT-based scores, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	5-27
5-23	Third-grade mathematics assessment sample, ECLS-K:2011 spring third-grade data collection: School year 2013–14	5-28
5-24	Fourth-grade mathematics assessment sample, ECLS-K:2011 spring fourth-grade data collection: Spring 2015	5-29
5-25	Fifth-grade mathematics assessment sample, ECLS-K:2011 spring fifth-grade data collection: Spring 2016	5-30
5-26	Mathematics assessment statistics by IRT-based score, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	5-31
5-27	Math assessment theta statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	5-32

List of Tables—Continued

Table		Page
5-28	Math assessment scale score statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	5-34
5-29	Mathematics assessment reliabilities, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	5-36
5-30	Third-grade science assessment sample, ECLS-K:2011 spring third-grade data collection: Spring 2014	5-37
5-31	Fourth-grade science assessment sample, ECLS-K:2011 spring fourth-grade data collection: Spring 2015	5-38
5-32	Fifth-grade science assessment sample, ECLS-K:2011 spring fifth-grade data collection: Spring 2016	5-39
5-33	Science assessment statistics by IRT-based score, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016	5-40
5-34	Science assessment theta statistics, by child characteristics, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016	5-41
5-35	Science assessment scale score statistics, by child characteristics, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016	5-43
5-36	Science assessment reliabilities, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, third-grade, fourth-grade, and fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016	5-45

List of Tables—Continued

Table		Page
5-37	Correlations of IRT theta score across rounds, by subject, ECLS-K:2011: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016	5-51
5-38	Correlations of IRT theta score across subjects, by round, ECLS-K:2011: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	5-53
6-1	Dimensional Change Card Sort variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School year 2012–13, spring 2014, spring 2015, and spring 2016	6-9
6-2	Mean Dimensional Change Card Sort total score for spring second grade, spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2013, spring 2014, spring 2015, and spring 2016	6-10
6-3	Numbers Reversed variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016	6-22
6-4	Mean Numbers Reversed *W*-ability score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	6-24
6-5	Mean Numbers Reversed age standard score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	6-25
6-6	Mean Numbers Reversed grade standard score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	6-26
6-7	Mean Numbers Reversed age percentile score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	6-27

List of Tables—Continued

Table		Page
6-8	Mean Numbers Reversed grade percentile score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	6-28
6-9	Flanker variable names, descriptions, value ranges, weighted means, and standard deviations for spring fourth grade and spring fifth grade: Spring 2015 and spring 2016	6-37
6-10	Mean Flanker overall score for spring fourth grade and spring fifth grade, by data collection round and child characteristics: Spring 2015 and spring 2016	6-38
6-11	Correlations of executive function scores across rounds, by task: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016	6-41
6-12	Correlations of executive function scores across tasks, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016	6-42
6-13	Variance estimates in two-level models for Numbers Reversed *W*-ability scores and student-level covariates, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016	6-45
6-14	Variance estimates in three-level models for Numbers Reversed *W*-ability scores and student-level covariates, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016	6-46
6-15	Variance estimates in two-level models for DCCS scores and student-level covariates, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016	6-47
6-16	Variance estimates in three-level models for DCCS and student-level covariates, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016	6-48

List of Tables—Continued

Table		Page
6-17	Variance estimates in two-level models for Flanker scores and student-level covariates, by data collection round: Spring 2015 and spring 2016	6-49
6-18	Variance estimates in three-level model for Flanker scores and student-level covariates, by data collection round: Spring 2015 and spring 2016	6-49
6-19	Skew of Numbers Reversed, DCCS, and Flanker scores	6-50
8-1	Teacher-reported social skills scales variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	8-4
8-2	Teacher-reported social skills scales reliability estimates for fall and spring kindergarten, fall and spring first grade, and fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016	8-6
8-3	Within-round correlations of teacher-reported social skills for spring third grade, spring fourth grade, and spring fifth grade: Spring 2014, spring 2015, and spring 2016	8-9
8-4	Mean teacher-reported Self-Control score, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	8-10
8-5	Mean teacher-reported Interpersonal Skills score, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	8-11
8-6	Mean teacher-reported Externalizing Problem Behaviors score, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	8-12
8-7	Mean teacher-reported Internalizing Problem Behaviors score, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	8-13

List of Tables—Continued

Table		Page
8-8	Teacher-reported Approaches to Learning scale variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	8-15
8-9	Within-round correlations of the teacher-reported Approaches to Learning scale and the teacher-reported social skills scales for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016	8-16
8-10	Mean teacher-reported Approaches to Learning scale scores, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	8-17
8-11	Children's Behavior Questionnaire variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten and spring first grade: School year 2010–11 and spring 2012	8-20
8-12	Temperament in Middle Childhood Questionnaire variable names, descriptions, value ranges, weighted means, and standard deviations for spring second grade, spring third grade, spring fourth grade, and spring fifth grade: Spring 2013, spring 2014, spring 2015, and spring 2016	8-20
8-13	Reliability estimates for the teacher-reported Attentional Focus and Inhibitory Control scales for fall and spring kindergarten, spring first grade, spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School year 2010–11, spring 2012, spring 2013, spring 2014, spring 2015, and spring 2016	8-22
8-14	Intercorrelations among Attentional Focusing and Inhibitory Control scales for spring third grade, spring fourth grade, and spring fifth grade: Spring 2014, spring 2015, and spring 2016	8-23
8-15	Mean Temperament in Middle Childhood Questionnaire Attentional Focus scores, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	8-23

List of Tables—Continued

Table		Page
8-16	Mean Temperament in Middle Childhood Questionnaire Inhibitory Control scores, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016	8-24
8-17	Student-Teacher Relationship Scale variable names, descriptions, value ranges, weighted means, and standard deviations for spring kindergarten, spring first grade, spring second grade, and spring third grade: Spring 2011, spring 2012, spring 2013, and spring 2014	8-25
8-18	Reliability estimates for the Student-Teacher Relationship Scale teacher-reported Closeness and Conflict scores for spring kindergarten, spring first grade, spring second grade, and spring third grade: Spring 2011, spring 2012, spring 2013, and spring 2014	8-26
8-19	Mean Closeness and Conflict scores from the Student-Teacher Relationship Scale for spring third grade, by child characteristics: Spring 2014	8-27
8-20	Teacher-reported Working Memory variable names, descriptions, value ranges, weighted means, and standard deviations for spring third grade: Spring 2014	8-29
8-21	Reliability estimates for the teacher-reported Working Memory scores for spring third grade: Spring 2014	8-29
8-22	Mean teacher-reported Working Memory scale for spring third grade, by child characteristics: Spring 2014	8-30
8-23	Parent-reported Working Memory variable names, descriptions, value ranges, weighted means, and standard deviations for spring third grade and spring fourth grade: Spring 2014 and spring 2015	8-37
8-24	Reliability estimates for the parent-reported Working Memory scores for spring third grade and spring fourth grade: Spring 2014 and spring 2015	8-37
8-25	Mean parent-reported Working Memory scale for spring third grade and spring fourth grade, by child characteristics: Spring 2014 and spring 2015	8-38

List of Figures

Figure		Page
3-1	Three-parameter IRT logistic function for a hypothetical dichotomous test item	3-7
3-2	Three-parameter IRT logistic functions for seven hypothetical dichotomous test items with different difficulty (b)	3-8
3-3	Three-parameter IRT logistic functions for two hypothetical dichotomous test items with different discrimination (a)	3-9
3-4	Three-parameter IRT logistic function for a hypothetical polytomous test item of a two-item testlet	3-11
3-5	Example of an empirical dichotomous item characteristic curve (ICC) for a well-functioning item: ECLS-K:2011 fall 2009 field test	3-16
3-6	Example of an empirical dichotomous item characteristic curve (ICC) for a poorly functioning item: ECLS-K:2011 fall 2009 field test	3-17
3-7	Item characteristic curve (ICC) compared to item information function (IIF)	3-19
3-8	Example test information function (TIF)	3-22

List of Exhibits

Exhibit		
1-1	Data collection schedule: School years 2010–11 through 2015–16	1-3
2-1	Principal sources of Reading, Mathematics, and Science content specifications for the ECLS-K:2011 third-grade, fourth-grade, and fifth-grade assessments: Spring 2014, 2015, and 2016	2-16
5-1	Routing path for the direct child assessment in the ECLS-K:2011 third-grade year	5-3
5-2	Routing path for the direct child assessment in the ECLS-K:2011 fourth-grade year	5-3
5-3	Routing path for the direct child assessment in the ECLS-K:2011 fifth-grade year	5-4

List of Exhibits—Continued

Exhibit		Page
6-1	Data flag description for the computerized Dimensional Change Card Sort for fall and spring second grade, spring fourth grade, and spring fifth grade: School year 2012–13, spring 2015, and spring 2016	6-13
6-2	Data flag description for the computerized Dimensional Change Card Sort for spring third grade: Spring 2014	6-14
6-3	Data flag description for the Flanker for the spring of fourth grade and spring of fifth grade: Spring 2015 and spring 2016	6-39
7-1	Child questionnaire content and item-level variables for spring third grade, spring fourth grade, and spring fifth grade: Spring 2014, spring 2015, and spring 2016	7-3
8-1	Teacher-reported item-level variables on peer relationships in spring second grade, spring third grade, spring fourth grade, and spring fifth grade: Spring 2013, spring 2014, spring 2015, and spring 2016	8-32
8-2	Teacher-reported item-level variables on school liking in spring fourth grade and spring fifth grade: Spring 2015 and spring 2016	8-34
8-3	Parent-reported item-level variables on peer relationships and friendships in spring second grade, spring third grade, spring fourth grade, and spring fifth grade: Spring 2013, spring 2014, spring 2015, and spring 2016	8-39
8-4	Parent-reported item-level variables on school avoidance in spring fourth grade and spring fifth grade: Spring 2015 and spring 2016	8-41

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 children participating in the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) in the third-, fourth-, and fifth-grade data collections.[1] The focus of this volume is the seventh through ninth rounds of data collection: the spring 2014 third-grade, the spring 2015 fourth-grade, and the spring 2016 fifth-grade rounds. Readers interested in the earlier rounds should refer to the prior user's manuals and psychometric reports, as shown here:

- *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. 2015a), hereinafter referred to as the base-year User's Manual, for information about the general study methodology and the kindergarten rounds of data collection;
- *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) (Tourangeau et al. 2015b);
- *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES 2017-285) (Tourangeau et al. 2017)*;*
- *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a); and
- *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First-Grade and Second-Grade Psychometric Report. (*NCES 2018-183) (Najarian et al. 2018b).

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 included interviews with parents; self-administered questionnaires completed by teachers and school administrators; one-on-

[1] Although the study refers to 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), not all study children were in the modal grade due to retention in a grade or promotion to a higher grade ahead of schedule.

one assessments of children; and beginning in third grade, a computer-assisted self-administered questionnaire for children. During the kindergarten year, it also included self-administered questionnaires for nonparental before- and after-school care providers. The ECLS-K:2011 is sponsored by the National Center for Education Statistics (NCES) within the Institute of Education Sciences (IES) of the U.S. Department of Education.

The ECLS-K:2011 is the third and most recent study in the Early Childhood Longitudinal Study (ECLS) program, which comprises three longitudinal studies of young children: the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K); the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B); and the ECLS-K:2011. The ECLS program is unprecedented in its scope and coverage of child development, early learning, and school progress. It draws together information from multiple sources, including 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 recent 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 the ECLS 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 Third-Grade, Fourth-Grade, and Fifth-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 18,174 kindergartners enrolled in 968 schools across the United States.[2] Up until the second-grade year, there were two data collections: one at the beginning (fall) and one near the end (spring) of the school year. Beginning in third grade, data collection occurred only once, each spring. Exhibit 1-1 presents an overview of the data collection schedule.

Exhibit 1-1. Data collection schedule: School years 2010–11 through 2015–16

School year	Grade[1]	Data collections[2]
2010–11	Kindergarten	Fall 2010
		Spring 2011
2011–12	First grade	Fall 2011
		Spring 2012
2012–13	Second grade	Fall 2012
		Spring 2013
2013–14	Third grade	Spring 2014
2014–15	Fourth grade	Spring 2015
2015–16	Fifth grade	Spring 2016

[1] Grade indicates the modal grade for children who were in kindergarten in the 2010–11 school year. After the kindergarten rounds of data collection, children were included in data collection regardless of their grade level.
[2] All but two rounds of data collection include the entire sample of children. The fall first-grade data collection included approximately one-third of the total ECLS-K:2011 sample of children. The fall second-grade data collection included the same subsample selected for the fall of 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).

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- through fifth-grade data collections. However, an eligible student did not have to participate in one grade to be eligible for data

[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.
[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.

collection in the next grade. The fall data collections that were conducted in first and second grade were not repeated in third grade and beyond.[4] The sample for the third-grade, fourth-grade, and fifth-grade spring data collections included all 18,174 base-year respondents.

Students who were initially included in the third-, fourth-, and fifth-grade samples were eligible for actual data collection if they were living in the United States at the time of data collection. However, because of 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, 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.

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 had critical implications for the design of the ECLS-K:2011. Data were collected on a wide array of topics at a broad level rather than on a select set of topics in more depth. Additionally, several different people from different contexts in the study child's life were asked to provide information about the child. Therefore, the design of the study included the collection of information from the students, their parents/guardians, their teachers, and their schools. As noted earlier, during the kindergarten data collection, the study also collected information from children's before- and after-school providers.

Students were directly assessed in each round of the third-, fourth-, and fifth-grade years. The untimed assessments were administered to the sampled students, one-on-one, by a trained assessor. The students were assessed in spring in reading, mathematics, and science, as well as executive function.[5] In addition to the cognitive components, the direct assessments in these rounds included measurements of height and weight for all students in the spring of each year. Beginning in the third-grade round, students completed a self-administered questionnaire using audio computer-assisted self-interview (ACASI)

[4] The fall first-grade and fall second-grade data collections were conducted with a 30 percent subsample, rather than the full sample.
[5] 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.

technology, a software system that read the instructions and questionnaire items to the child, while the same text was displayed on a laptop's screen. Children chose answers to the questions by selecting responses directly on the touch-sensitive screen of the laptop. The questionnaire, which included varying content in the third-, fourth-, and fifth-grade rounds, covered such topics as relationships with peers and peer support; occurrences of peer victimization; and overall happiness with different aspects of life (e.g., attention from parents, hobbies and free time activities). Because some of the questions may have been deemed sensitive, the questionnaire was only administered once children were deemed old enough to be able to complete the questionnaire independently. In second grade, the National Institute on Deafness and Other Communication Disorders (NIDCD) provided funds to evaluate the hearing of the subsample of students participating in the fall data collection. This hearing evaluation was conducted again in the spring third- and spring fifth-grade rounds with the same students who were evaluated in the fall second-grade round.

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 using computer-assisted interviews (CAIs). Most parent interviews were conducted by telephone.[6] In the data collections conducted in the spring of 2014, spring of 2015, and spring of 2016, the parent interviews asked parents to provide demographic information, where the child was born (if missing from a previous round), and describe parental involvement in school, family structure, the child's specific ethnic origin, family literacy practices, the child's sleep patterns, field trips focused on science activities, household composition, parent depression, parent health, household food security, household income, parent education level, parent employment, parent military service, nonresident parents, and where the nonresident parent was born (if missing from a previous round). Parents were also asked to report on their children's health and disability status. In addition to these topics, the spring 2014, 2015, and 2016 parent interviews included the topics described below, summarized by the round of the study:

- In the spring of third grade, the parent interview included questions from previous rounds about where the parent was born (if missing from a previous round); whether a language other than English was spoken in the home; before-and after- school care; parental warmth; discipline; and communication with the child. Parents also reported on their children's experiences with peer victimization, as well as children's physical activity. New to the third-grade data collection were questions about whether the parent monitored that the child's homework had been completed; the child's specific ethnic origin (that is, to what specific Hispanic, Asian, or Pacific Islander group the child belonged, such as Puerto Rican, Chinese, Samoan, etc.); how many hours of sleep the child got on weeknights; and whether parents had been (since the child was born) or currently were on active duty in the military; questions to assess the child's working

[6] 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.

memory; and questions about whether the child had been on field trips focused on science activities.

- In the spring of fourth grade, the parent interview included questions from previous rounds about neighborhood safety; whether and how long the child took care of himself or herself before or after school; discipline; communication with the child; and welfare and other public transfers. Parents also reported on their children's working memory and children's physical activity. New to the fourth-grade data collection were questions about parents' use of a computer or other electronic device to find out about children's homework; school assignments, grades, and how children at the school were doing as a group; parent reports of the child's grades; the frequency that the child avoided school; family monitoring of what the child looked at online and how many hours were spent online; children's friendships; how frequently the parent and child argued; and overall life stress in the past 12 months.

- In the spring of fifth grade, the parent interview included questions from previous rounds about whether a language other than English was spoken in the home; school avoidance; the home environment; marital/partner satisfaction; children's friendships; child care before or after school and self-care; discipline and emotional supportiveness; and welfare and other public transfers. Parents also reported on their children's physical activity, dental care, and routine health care.

  The fifth-grade parent interview mostly contained questions that had been asked in at least one round of the study. However, several questions about animals and their use to help children with disabilities were added to the child's health and well-being section. Also, several questions from earlier rounds of the study that had not been fielded in recent rounds were included in order to have a final data point in the study (e.g., parent's educational expectations for the child, marital/partner satisfaction, use of a language other than English, and outings with the child). Questions that were new to the fourth-grade data collection were retained in the fifth grade (questions about parents' use of a computer or other electronic device to find out about children's homework, school assignments, grades, and how children at the school were doing as a group; parent reports of the child's grades; the frequency that the child avoided school; family monitoring of what the child looked at online and how many hours were spent online; children's friendships; how frequently the parent and child argued; and overall life stress in the past 12 months).

Teachers in third grade, as in earlier rounds, 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.

Beginning in fourth grade, the ECLS-K:2011 changed its approach to collecting teacher questionnaire data to account for the fact that as children moved into the upper elementary grades, more than one teacher was often involved in a given child's instruction. In order to accommodate the variation in organization, the spring 2015 fourth-grade and spring 2016 fifth-grade data collections identified each

study child's reading teacher, and that teacher was asked to complete questionnaires that the general classroom teacher previously completed. In addition, science and mathematics teachers were also identified. To reduce burden on teachers, half of the sampled children were randomly assigned to have their mathematics teacher complete questionnaires, while the other half of the sampled children were randomly assigned to have their science teacher complete questionnaires. Thus, every child had a reading and either a mathematics or science teacher identified. All identified teachers received a self-administered teacher-level questionnaire. This questionnaire asked about time allocated to various subjects and activities for students in the given grade and about the teacher's evaluation methods, attitudes about school climate, educational background, and teacher experience.

Three additional subject-specific child-level questionnaires were also distributed for the identified reading and mathematics or science teacher. Each of these questionnaires included child-level questions and classroom questions. Further details on the content of the various questionnaires and how they differed across rounds can be found in the relevant years' user's manuals listed earlier.

Special education teachers and service providers of sampled students who had an Individualized Education Program (IEP) in third, fourth, or fifth 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 third-, fourth-, and fifth-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 third-, fourth-, and fifth-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 spring of 2014 (third grade), 2015 (fourth grade), and 2016 (fifth grade). 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 third-, fourth-, and fifth-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 third, fourth, and fifth grades. Chapter 7 provides information on the Child Questionnaire and the different constructs measured. Chapter 8 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 8 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 are prepared for and used in PARSCALE, the computer program used for estimating item response theory (IRT) models from which assessment scores are produced, as well as what quality control checks are performed on the assessment data. Appendix C lists the ECLS-K:2011 third-, fourth-, and fifth-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 skills and knowledge 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 were expected to be in fifth grade), thus allowing for comparisons of achievement across grades and to quantify the gains children made 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) to allow for cross-cohort comparability in the knowledge and skills measured in each study. Despite this overlap, the direct cognitive assessment scores released for the ECLS-K:2011 in the kindergarten, kindergarten through first grade, kindergarten through second grade, kindergarten through third grade, kindergarten through fourth grade, and kindergarten through fifth grade data files are not directly comparable to those originally delivered for the ECLS-K.[1] However, currently in progress is the development of re-scaled scores from the ECLS-K assessment that will be comparable with the ECLS-K:2011. These re-scaled ECLS-K scores will be released as a separate dataset and will be directly comparable with the ECLS-K:2011 scores provided on the kindergarten through fifth grade data file. The ECLS-K scores had to be rescaled to reflect that the ECLS-K:2011 assessment included 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 ECLS-K:2011 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.

[1] For the ECLS-K:2011, a longitudinal, restricted-use data file was created after each grade, and a longitudinal, public-use data file was created after most grades. The public-use files are as follows: ECLS-K:2011 Public-Use Kindergarten–First Grade Data File and Electronic Codebook (NCES 2015-086); ECLS-K:2011 Public-Use Kindergarten–Second Grade Data File and Electronic Codebook (NCES 2017-286); ECLS-K:2011 Public-Use Kindergarten–Fourth Grade Data File and Electronic Codebook (NCES 2018-033); ECLS-K:2011 Public-Use Kindergarten–Fifth Grade Data File and Electronic Codebook (NCES 2019-050).

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 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 fifth 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 specifications for third, fourth, and fifth grades are based on the fourth- and eighth-grade National Assessment of Educational Progress (NAEP) Reading Frameworks for 2011 (National Assessment Governing Board 2010), current curriculum standards from Texas, California, New Jersey, Florida, and Virginia,[3] and the Common Core State Standards (see exhibit 2-1 for source information). The reading specifications were reviewed by content staff at the Educational Testing Service (ETS) and were drawn from a composite framework based on these states. In addition, the current framework specifications are built upon those developed for the earlier grades (kindergarten, first grade, and second grade), which were largely based on the 2009 NAEP Framework (National Assessment Governing Board 2008), the ECLS-K kindergarten, first-grade, third-grade, and fifth-grade reading assessment frameworks,[4] in addition to standards from the same five states for earlier grades. 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 third-grade, fourth-grade, and fifth-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 fifth grade is included here as a reference for what content was targeted in the prior rounds as well as how the third-grade, fourth-grade, and fifth-grade content relates both to the overall framework and to the content measured in earlier rounds.

[3] Curriculum standards from Texas, California, New Jersey, Florida, and Virginia were selected to inform development of the third- through fifth-grade reading specifications providing continuity with the kindergarten through second-grade reading specifications, as these same states were utilized in that framework design. Specifications from these states were selected by ETS content staff as representative of typical U.S. state standards.

[4] There were no frameworks or national administrations of second-grade and fourth-grade assessments in the ECLS-K.

2.1.1.1 Basic Reading Skills

Basic reading skills include many early literacy skills in subcategories such as phonological awareness, familiarity with print, recognition of letters and sounds, and identification of common sight words. Although all of the subcategories of basic reading skills are described below for completeness, items from only the subcategory sight words were administered in the third-, fourth-, and fifth-grade assessments. The difficulty levels of items from the other subcategories were too easy for the later grade assessments, but were appropriate for use at the earlier grade levels.

Phonological awareness is one of the reading skills included in the ECLS-K:2011 assessments at the early grades. 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 assessments in the early grades measure the following types of phonological awareness skills:

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

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

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 asked children to perform tasks such as choosing a specific letter from a set or giving the name of a letter that was shown to them. These tasks involve identification of both upper and lower case letters. In addition, children were asked to associate a letter with its sound. These tasks include the child identifying the letter that makes a sound vocalized by the assessor or the child vocalizing the sound represented by a certain letter named by the assessor.

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

2.1.1.2 Vocabulary

Vocabulary knowledge represents understanding of the meanings of words. Although children may be able to decode printed text, they also must understand the meaning of the words they have read in order to be able to comprehend the text. Vocabulary test questions in the ECLS-K:2011 assessment ask children to convey their vocabulary knowledge both verbally (expressive vocabulary) and nonverbally (receptive vocabulary). To test 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 the image representing that word from a group of similar images, nonverbally (e.g., pointing to a picture presented in the assessment easel). This task assesses the child's understanding of a given word in relation to a picture of it. In addition, some

receptive vocabulary tasks include 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 third-, fourth-, and fifth-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 at these grades.

2.1.1.3 Comprehension

As noted earlier, the ECLS-K:2011 reading framework was modeled after the NAEP 2011 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 2010).

- **Locate/recall.** Assessment items 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 identify the main idea of a passage.

- **Integrate/interpret.** Assessment items 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 to provide a difference between the descriptions of two bedrooms in a passage.

- **Critique/evaluate.** Assessment items 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 to identify the most likely reason the author wrote a particular passage.

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, third, fourth, and fifth grades, 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

One of the goals of the ECLS-K:2011 is to enable cross-cohort comparisons. Consequently, continuity between the ECLS-K (1998–99) and ECLS-K:2011 framework specifications was necessary in order 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 2000), 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

Content covered in third, fourth, and fifth grades was determined by comparing the state or national standards from Texas, Virginia, NAEP, and the National Council of Teachers of Mathematics (NCTM) (see exhibit 2-1 for source information.) Common Core State Standards were not used in the comparison since these standards are similar to the national standards set by NCTM and NAEP. As in reading, the framework for third, fourth, and fifth grades builds on the framework developed in kindergarten through second grade, using the same sources.[5] The target content percentages for third, fourth, and fifth grades are based on the fourth-grade recommendations in the 2011 NAEP frameworks. The numbers for the national assessments maintain congruency in the primary (kindergarten through second grade) and intermediate (third, fourth, and fifth grade) elementary grade bands. The framework

[5] The standards of other states (California, New Jersey, and Tennessee) were also used in development of the kindergarten through second grade framework. However, these states have now adopted the Common Core State Standards, and thus, have been excluded due to their similarities to NCTM and NAEP.

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**.

2.1.2.1 Number Properties and Operations

From kindergarten through fifth grade, the Number Properties and Operations 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 third-, fourth-, and fifth-grade assessments, this content category is measured with questions assessing 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 third-, fourth-, and fifth-grade assessments, this content category is measured with questions asking children to estimate the weight of an object on a scale, perform basic operations that require knowledge of money, and compute area.

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 third-, fourth-, and fifth-grade assessments, this content category is measured with questions asking children to combine basic shapes into other basic shapes and the measurement of the diameter and circumference of a circle.

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 third-, fourth-, and fifth-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 are added. In the third-, fourth-, and fifth-grade assessments, this content category is measured with questions asking children to complete sequences and to solve simple equations.

2.1.3 Science Test Specifications

For the ECLS-K:2011 framework development for third, fourth, and fifth grades, the standards of six states (Arizona, California, Florida, New Mexico, Texas, and Virginia) in addition to those from the 2009 NAEP standards were analyzed to find a commonality of topics that are taught at each grade level (see exhibit 2-1 for source information). Three to four standards were designated for each grade level in each of four reporting categories: **scientific inquiry,**[6] **life science**, **physical science,** and **Earth/space science**. Standards from the same states were used in developing the kindergarten through second-grade framework and were also drawn upon for the third- through fifth-grade specifications in order to permit alignment of the standards across grade levels.

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 third-, fourth-, and fifth-grade assessments are items about independent variables, interpretation of data from observations, and reading a contour map.

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 subsequent grades, children are expected to identify the three states of matter (solid, liquid, and gas); 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 third-, fourth-, and fifth-grade assessments,

[6] The National Science Education Standards: (1) Science as Inquiry; (5) Science and Technology; (6) Personal and Social Perspectives; (7) History of Nature and Science; and (8) Unifying Concepts and Processes in Science were collapsed into the single category, Scientific Inquiry, for the ECLS-K:2011 (National Research Council 1996).

this content category is measured with questions about solar energy, the materials from which common objects are made, and decomposition.

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 third-, fourth-, and fifth-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 third-, fourth-, and fifth-grade assessments, this content category is measured with questions about shadows, the solar system, and renewable resources.

2.2 Item and Time Allocation Across and Within Subject Areas

For the third-, fourth-, and fifth-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 assessments 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 fifth 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-, third-, fourth-, and fifth-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 fifth-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	†	20	30	35	15
4	†	20	30	35	15
5	†	15	25	40	20

† Not applicable.
NOTE: Basic skills and Vocabulary percentages are combined into the Vocabulary content category.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), cognitive assessment frameworks, fall 2008 and fall 2012.

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 that continues into fourth and fifth grades 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 fifth-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
4	40	20	15	10	15
5	30	18	18	12	22

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), cognitive assessment frameworks, fall 2008 and fall 2012.

[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 fifth-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
4	25	25	25	25
5	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), cognitive assessment frameworks, fall 2008 and fall 2012.

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. 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 was 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.[8]) 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[9] 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. It was not fully adaptive since each child's routing to subsequent individual assessment items was not individually determined by responses to a prior item.

[8] Details on the SERS are included in the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11, First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b).

[9] 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 of the ECLS-K:2011 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* (NCES 2018-182) (Najarian et al. 2018a). A single-stage Spanish early reading skills assessment[10] 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 needed 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

[10] 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.

administered in English. However, several of the assessments were translated into and administered in Spanish in the 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 subsequent 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, and the development of these assessments can be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b).

Exhibit 2-1. Principal sources of Reading, Mathematics, and Science content specifications for the ECLS-K:2011 third-grade, fourth-grade, and fifth-grade assessments: Spring 2014, 2015, and 2016

Domain and sources
Reading
Reading Frameworks for fourth and eighth grade: 2011 National Assessment of Educational Progress: http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/reading-2011-framework.pdf National Assessment Governing Board (2000). *Reading Framework for the National Assessment of Educational Progress: 1992–2000*. Washington, DC: Author.
Texas curriculum standards: http://ritter.tea.state.tx.us/rules/tac/chapter110/index.html
California curriculum standards: http://www.cde.ca.gov/be/st/ss/documents/elacontentstnds.pdf
New Jersey curriculum standards: http://www.nj.gov/education/cccs/2004/s3_lal.pdf
Florida curriculum standards: http://www.cpalms.org/Public/search/Standard
Virginia curriculum standards: http://www.doe.virginia.gov/testing/sol/standards_docs/english/2010/stds_all_english.pdf
Common Core State Standards: http://www.corestandards.org/ELA-Literacy/
Mathematics
Texas Response to Curriculum Focal Point for Kindergarten through Grade 8 Mathematics: http://txar.org/docs/txcfps_final_2_1_10.pdf.
Virginia Department of Education Standards of Learning and Testing, Mathematics: http://www.doe.virginia.gov/testing/sol/standards_docs/mathematics/index.shtml

See notes at end of exhibit.

Exhibit 2-1. Principal sources of Reading, Mathematics, and Science content specifications for the ECLS-K:2011 third-grade, fourth-grade, and fifth-grade assessments: Spring 2014, 2015, and 2016—Continued

Domain and sources
Mathematics—Continued
Mathematics Framework for the 2011 National Assessment of Educational Progress: https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/mathematics/2011-mathematics-framework.pdf
National Council of Teachers of Mathematics, Standards and Focal Points: https://www.nctm.org/Standards-and-Positions/Principles-and-Standards/Principles,-Standards,-and-Expectations/
Science
Arizona state science standards: https://www.azed.gov/standards-practices/k-12standards/standards-science/
California state science standards: https://www.cde.ca.gov/be/st/ss/documents/sciencestnd.pdf
Florida state science standards: http://www.fldoe.org/academics/standards/subject-areas/math-science/science/
New Mexico state science standards, https://webnew.ped.state.nm.us/wp-content/uploads/2017/12/MSB_contentstandards_CurrentScienceStandardsV2.pdf
Texas state science standards: http://ritter.tea.state.tx.us/rules/tac/chapter112/index.html
Virginia state science standards: http://www.doe.virginia.gov/testing/sol/standards_docs/science/2003/index.shtml
Science Framework for the 2009 National Assessment of Educational Progress: http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/science-09.pdf
The National Science Education Standards: http://www.csun.edu/science/ref/curriculum/reforms/nses/

NOTE: Links were active at the time the sources were accessed during the development of the content specifications.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011).

3. ANALYSIS METHODOLOGY

This chapter describes the standard procedures used to process data from the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (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 fifth-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, 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 multiple-choice items for which they responded "don't know" that were followed by at least one subsequent item the test taker did

[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.

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 Pearson product-moment 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 (Crocker and Algina 1986). 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., third to fourth 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.

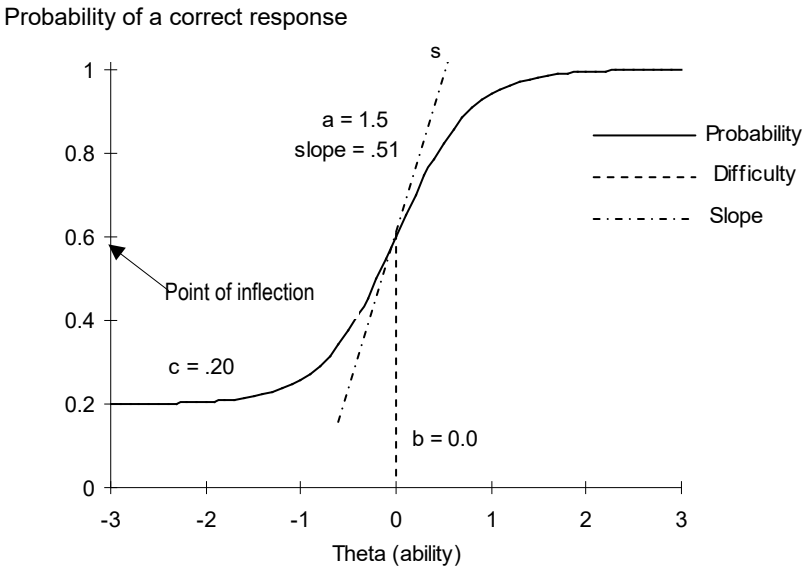
Figure 3-1 is an example of a graph of the logistic IRT function for a hypothetical dichotomous test item. The graph shows the most general model, the three-parameter (3PL) IRT model. The three item parameters are *a* (discrimination), *b* (difficulty), and *c* (guessing). The horizontal axis represents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

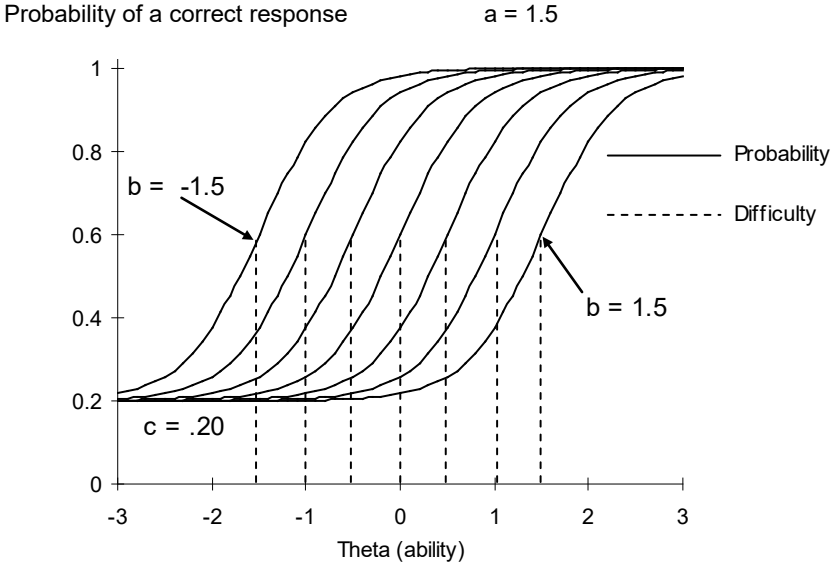
The IRT *c* parameter represents the probability that a test taker with very low ability will answer a multiple-choice item correctly. In figure 3-1, out of 100 people with very low ability, 20 would get the item correct. Note that the *c* parameter does not necessarily equal 1 divided by the number 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 ($-\infty$) 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*)

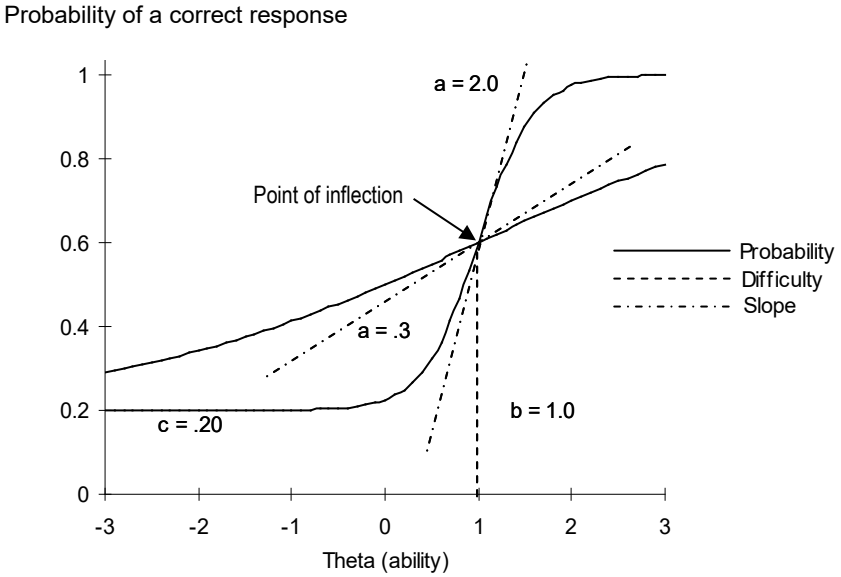


NOTE: IRT= item response theory. *a* = parameter for discrimination; *b* = parameter for difficulty; and *c* = guessing parameter.
SOURCE: U.S. Department of Education, National Center for Education Statistics, *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric 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 = $0.425a_i(1\text{-}c_i)$). 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: IRT= item response theory. 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.)

[2] High levels of dependency result in biased item parameter estimates, 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.

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 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 i^{th} item, a response that is scored in the j^{th} 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

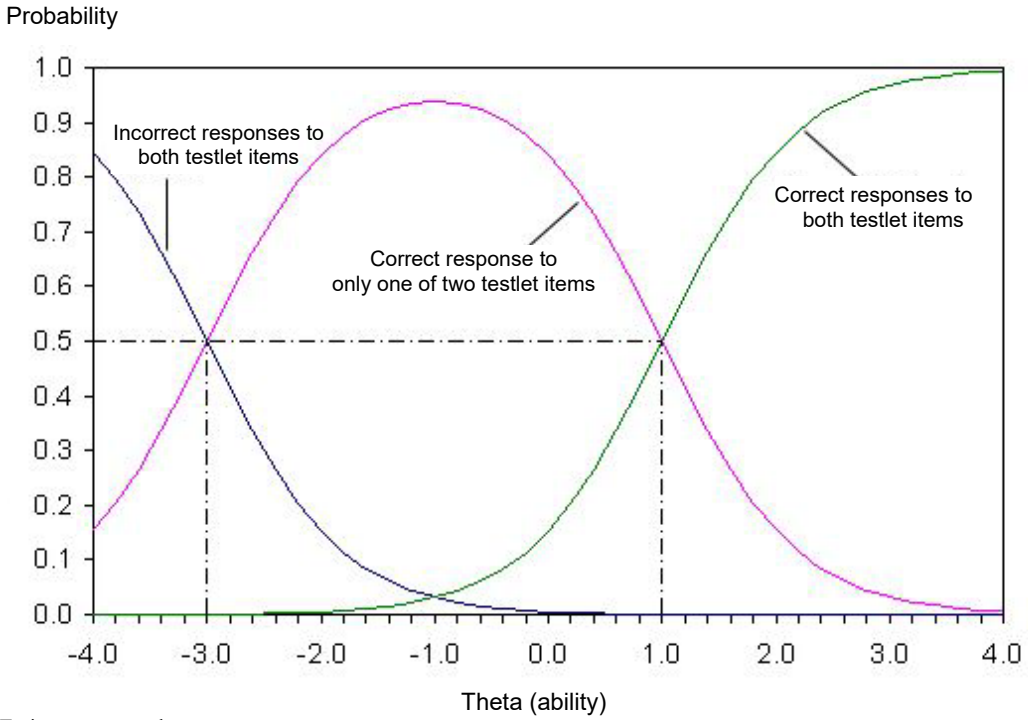
m_j = number of items in the testlet or associated with the same passage in polytomous item j;

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



NOTE: IRT= item response theory.
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[3] 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.

[3] Constant guessing is defined as fixing the guessing parameter for multiple choice items to 1/(number of options).

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.[4] 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. For the purpose of this analysis, 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-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.[5]

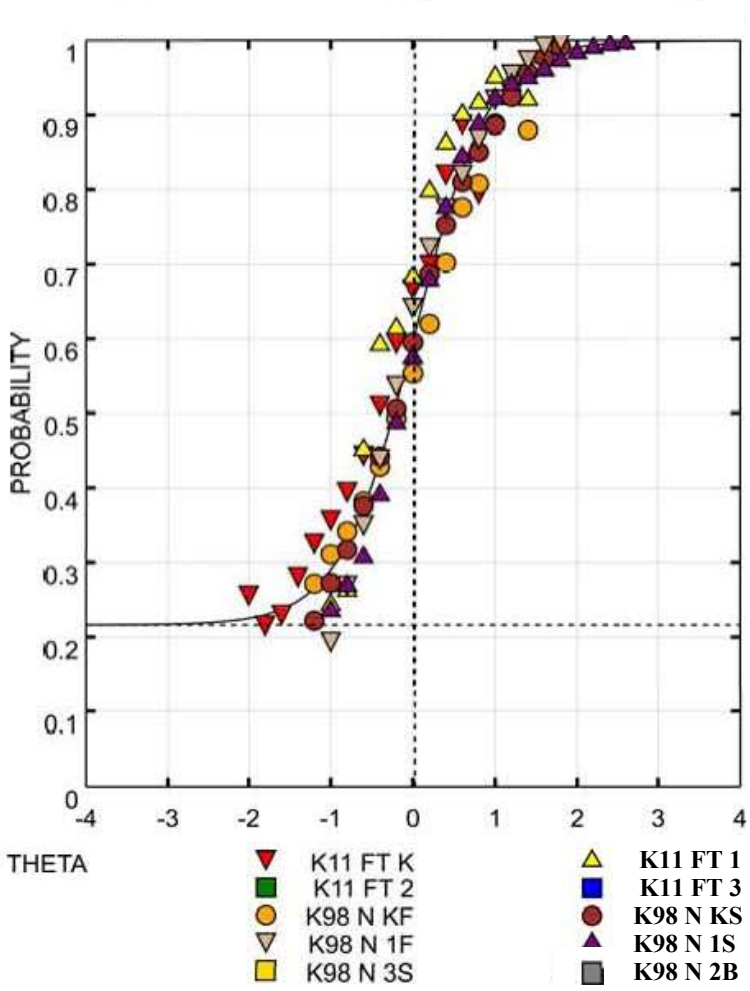
[4] In analyses of the national data, other factors, such as the item's content category, and where along the ability distribution the model discrepancies occurred, factored in to the decision to exclude an item from scaling and scoring.
[5] The same model used in the first round of the study (the fall of kindergarten) must be used in subsequent rounds to permit longitudinal measurement across the entire study.

3.2.5 Evaluating Items Using Empirical Item Characteristic Curves

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.

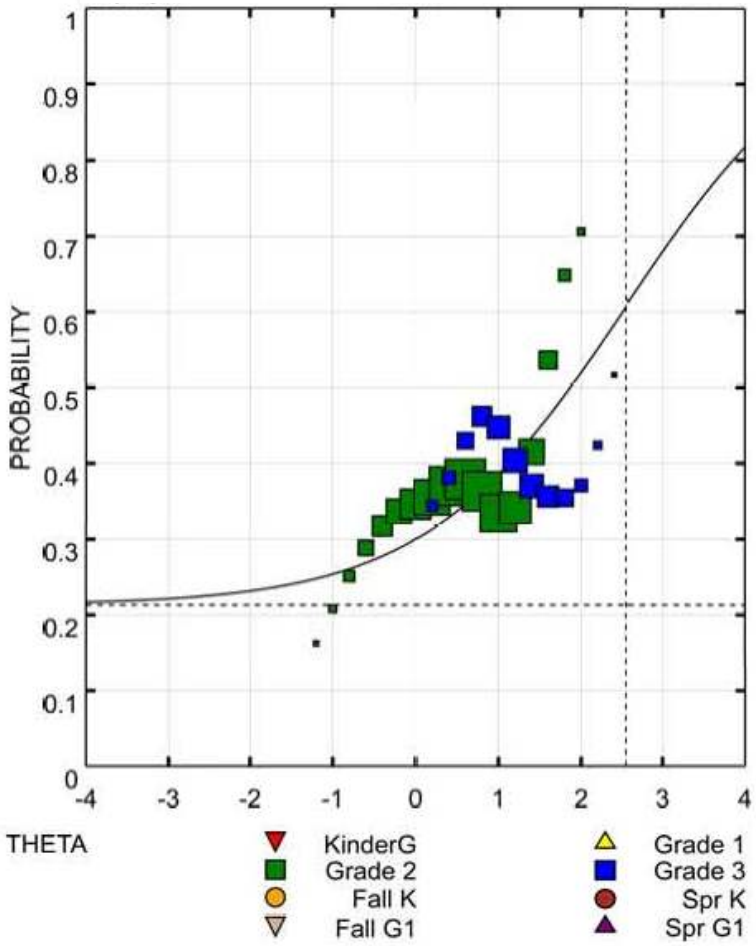
Figure 3-6 shows the empirical ICC of a dichotomous item included in the assessment field test that did not function particularly well. Although about 37 percent of the test takers answered this item correctly, performance on this item was not strongly related to overall ability; throughout most of the 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.

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

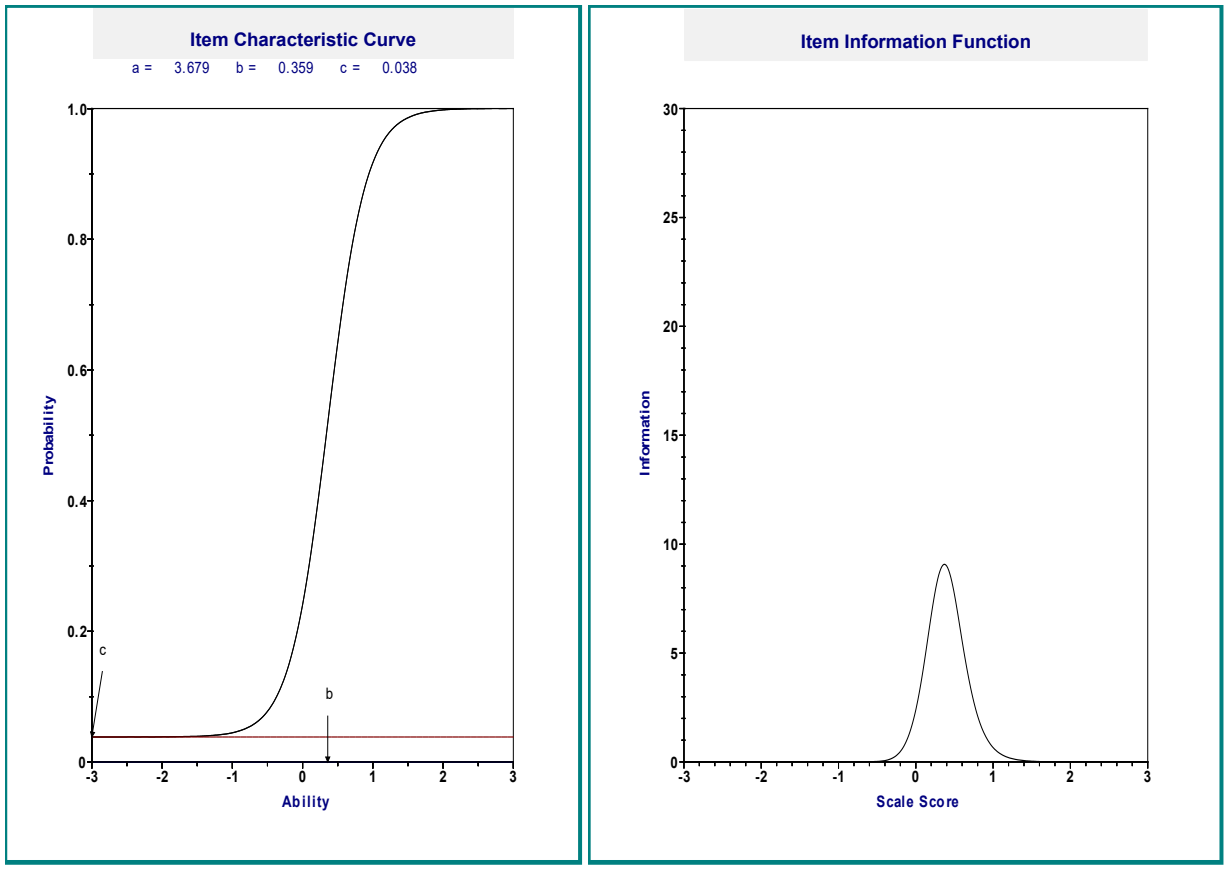
Another way to measure item and test quality is to estimate the item information function (IIF) and test information function (TIF). In psychometrics, the precision of parameter estimates at the various ability levels can be measured using the information function (Lord 1980). This is computed as a function 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

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

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

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:

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

where

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

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

θ = ability of the test taker;
j = j^{th} 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 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.[6]

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:

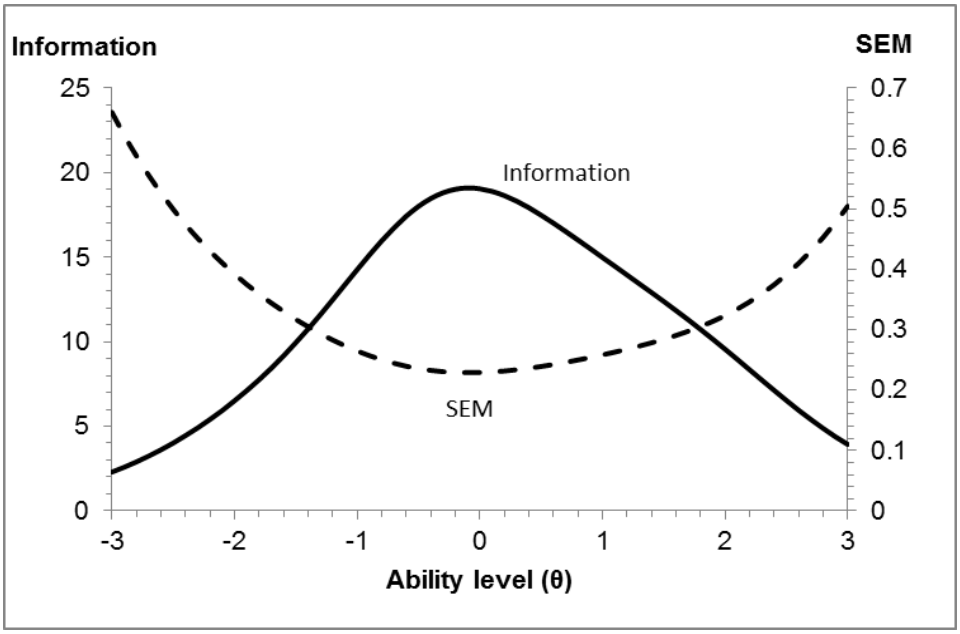
$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

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.

[6] 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.[7] 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 based on guidelines adopted from the National Assessment of Educational Progress (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

[7] 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.

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.[8] 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.[9] That is, after all of the kindergarten through fifth-grade assessments were completed, there were nine 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 spring of third grade is, on average, greater than the levels of performance in the spring of second grade but lower than the level of performance in the spring of fourth grade.

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 fall round for kindergarten, first grade, and second grade, where assessments in both fall and spring were administered) 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 recalibrating all of the item parameters using Bayesian priors[10] 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

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

[9] 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, the fall or spring of second grade, the spring of third grade, the spring of fourth grade, or the spring of fifth 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.

[10] 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.

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 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 reduced more in an effort to create more realistic estimates.

Theoretically, this approach has much to recommend it. In practice, the model has to have reasonable estimates (i.e., better estimation of outliers in the ability distributions) of the difference in ability levels among the subpopulations (different data collection rounds) to incorporate realistic Bayesian priors for the ability and item parameter estimates. The PARSCALE program generates initial item parameter estimates from default values or item difficulty statistics of a Bayesian prior calculation with a similar, or the same, population. Similarly, item parameter Bayesian priors and *a priori* distributions of abilities by subpopulation may be generated by PARSCALE or input from Bayesian prior distributions. Essentially, the within-grade longitudinal scales are determined by the items, and the initial Bayesian prior ability means for the children in the different rounds are in turn determined by the differential performance of the children on these items across rounds. The approach of using adaptive testing procedures combined with Bayesian procedures that allow for the use of prior values on both ability distributions and the item parameter estimates is needed in longitudinal studies to minimize 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 third-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 proportion 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, 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 fifth-grade reading item associated with a passage, if differential performance was shown by children matched on total score, it might be determined that the item and/or the passage required the reader to be familiar with certain cultural traditions that may favor a particular subgroup compared to others.

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 odds ratio is defined as follows (Zwick 2012):

$$\hat{\alpha}_{MH} = \frac{\sum_k N_{R1k} N_{F0k} / N_k}{\sum_k N_{R0k} N_{F1k} / N_k}.$$

In this equation, k is the number of score categories on the matching criterion, N_{R1k} indicates the number of test takers in the reference group who answered correctly, N_{F1k} indicates the number of test takers in the focal group who answered correctly, N_{R0k} indicates the number of test takers in the reference group who answered incorrectly, N_{F0k} indicates the number of test takers in the focal group who answered incorrectly, N_k is the total number of examinees, and α_{MH} is the population parameter that is assumed to be constant over all levels of the matching criterion.

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.[11] Accordingly, 41 two-way tables were produced for each item, one for each theta interval. The first dimension of each of the 41 two-way tables is population subgroup (e.g., White children versus Black children), and the other dimension is whether 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

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

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 sight words, while others attend schools where relatively more time is spent on developing reading comprehension. Both groups may have similar total scores in reading, but the reading comprehension 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 Other Pacific Islanders). There were too few non-Hispanic American Indian/Alaska Native and children of Two or more races 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 Fifth-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 fifth 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, the fall and spring of second grade, the spring of third grade, the spring of fourth grade, and the spring of fifth 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 fourth-grade low and the fourth-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 began in the spring of kindergarten since a science assessment in the fall of kindergarten was not administered.[12]

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 were 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 acquired 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 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 adjacent round 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 was contrasted. Items that were not common functioning would not be used as common items for the purposes of developing a vertical scale. These items were treated as completely different items in the calibration and scoring by round, unlike common items, which were 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

[12] 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.

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, second-grade, third-grade, fourth-grade, and fifth-grade assessment items are included in section 5.2.2.

3.5.2 Concurrent Calibration and Chain Linking

Once all rounds of data collection for the ECLS-K:2011 were complete, the methodology used to produce the final longitudinal scales and scores for each domain was reviewed to assure that the methods chosen at the beginning of the study were producing the most precise scores possible. As is typical in longitudinal scaling situations, there are a number of possible ways to use IRT to produce a scale. At one end of the spectrum are approaches that impose relatively few assumptions about the equality of item parameters across data collection rounds for common items during the scaling phase of analysis. These methods scale rounds (or grades, in the case of the ECLS-K:2011) of data separately and then link these separate IRT calibrations on a single common longitudinal scale by determining scale transformations using procedures like the Stocking and Lord (1983) approach that best match the test characteristic curves for the items common to adjacent grades. Such transformations are determined so that the expected average score on the common items, conditional on scale score, is equivalent across grades—a somewhat weaker assumption than requiring each common item to have identical item parameters. These methods, therefore, preference model-data fit at the cost of requiring large numbers of item parameters to be estimated. At the other end of the spectrum are approaches that impose stronger assumptions about the equality of IRT item parameters across multiple collection points through the estimation of a single concurrent IRT calibration based on a pooled dataset, and direct production of a set of results on a single longitudinal scale. Between these two extremes lie options that pool certain rounds of data, which produce fewer separate scalings that need to be linked. The following sections describe the methods reviewed and the resulting selection of methodology to produce the final longitudinal scales and scores.

3.5.2.1 Methodology Approaches

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 in the ECLS-K:2011 (which followed children from kindergarten in the 2010-11 school year through the spring of 2016).

As a starting point, ETS compared results from a largely unconstrained scaling approach (i.e., the grade-by-grade scaling and linking done to date) and a fully constrained scaling. The purpose of the analyses was to provide guidance as to whether either of these approaches, or some other possible variant, was the best way to produce the final scales and scores that precisely measure growth over time.

Prior rounds of the ECLS-K:2011 were scaled using an ***unconstrained*** approach that began by pooling and concurrently calibrating the data from the fall and spring rounds within a grade level in the kindergarten through second-grade rounds, where both fall and spring assessments were administered. (Third-, fourth-, and fifth-grade assessments were administered only in spring.) Then, the subsequent link between the assessments used at different grade levels relied on the presence of common items shared across the tests and was developed using chain-linking methodology described in Stocking and Lord (1983). The combination of concurrent calibration *within* grade (for the kindergarten through second-grade rounds), and chain linking *across* grades, resulted in a single, longitudinal scale.[13]

Within kindergarten through second grade, each of the rounds of data collection—fall and spring—was 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 were 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

[13] 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.

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. The unconstrained approach seeks to provide a best fit to each wave of data. However, that fit is purchased at a price: a large number of ICCs must be estimated (i.e., separate ICCs must be estimated for each item in each of the grades in which that item was administered); item-level sample sizes for estimating the ICCs (though substantial in most cases) are more modest than would be available if common functioning assumptions were imposed across grades; and multiple sets of linking parameters must be estimated, potentially introducing unnecessary estimation errors in the results.

As a point of comparison, a ***fully constrained*** approach was also evaluated. In this approach a single, concurrent calibration was run pooling all of the data collected (kindergarten through fifth grade). Separate means and standard deviations were estimated for each of the nine rounds of data collection (fall kindergarten, spring kindergarten, fall first grade, spring first grade, fall second grade, spring second grade, spring third grade, spring fourth grade, and spring fifth grade); and a single ICC was estimated for each common item regardless of the number of rounds in which each common item was administered. The constrained approach estimates far fewer item parameters and utilizes larger sample sizes in the estimation of the parameters for each item, enabling more precise estimates. It also does not require estimated linking parameters, though that is slightly offset by having to estimate proficiency means and standard deviations for each round.[14] The tradeoff is that such a model can result in somewhat poorer model-data fit due to strong equality constraints imposed on item parameter estimates.

A ***partially constrained*** approach, providing a balance between the fully constrained and unconstrained solutions, was developed by producing two separate calibrations: one with pooled data from fall kindergarten, spring kindergarten, fall first grade, spring first grade, fall second grade, and spring second grade; and one with pooled data from spring third grade, spring fourth grade, and spring fifth grade. Within each of these calibrations, separate ability (theta) distributions were estimated for each data collection round but a single common item characteristic curve was estimated for each item. The results of these separate scalings were then linked using the Stocking and Lord (1983) procedure to place the third-, fourth-, and fifth-grade calibration results on the kindergarten through second-grade scale. The advantage of this intermediary solution is that it seeks to address concerns about content comparability and score meaning over large grade spans (see, for example, Kolen and Brennan 2004, pages 412-413) without sacrificing too much of the parsimony associated with the fully constrained model. However, like the fully constrained model, some assumptions about equality of item parameters across data collection rounds are

[14] Note that, in the unconstrained solution the means and standard deviations for these grades are set at 0 and 1 *a priori* to resolve the linear indeterminacy in the IRT scale.

being imposed. Therefore, this approach can still exhibit somewhat greater model-data misfit than the fully unconstrained scaling.

3.5.2.2 Comparison of Approaches and Methodology Selection

Results from the three approaches were compared on several aspects. The goal of these comparisons was to determine the degree to which the more restrictive, parsimonious models produced results that were consistent with those from the less restrictive models. Since there is no known correct answer to evaluate the results against, the logic of the analysis approach was to prefer more restrictive, parsimonious models if the results they produced were similar to those from the unrestricted models. As sample sizes for such comparisons are quite large, formal statistical criteria are of less value in making the choice of which scaling approach to choose. Therefore, the evaluation was based largely on various graphical displays and descriptive indices that were reviewed by the contractor and by NCES staff.

Comparisons of the aggregate score distributions (in the θ -metric as well as the scale score metric) were carried out. The comparisons looked at the similarity of the patterns of average growth, the dispersion of scores, and the shapes of the score distributions within and across the cohorts/grades. Correlations between scores produced by the various methods were also examined to determine whether the different scaling approaches produced the same rank-ordering of test-takers within each grade and cohort.

After placing all results in a common metric, patterns of average growth were nearly identical. In the θ scores, there were some differences in the pattern of change for within-grade standard deviations for mathematics and reading. These standard deviations were reduced at a steeper rate in the fully constrained and partially constrained solutions than was the case for the unconstrained solution. The same was not observed for the science results. For all three subject areas, the scale score results show little evidence of any differences in standard deviations across scaling approaches. This latter finding (i.e., that differences apparent in the θ scores are not evident in the scale scores) is due to the fact that different estimates of latent variable models (i.e., the results in the θ -metric) often make nearly identical predictions about observable features of the data.[15] The shapes of the distributions in mathematics and science were quite similar in both the scale score and θ -metrics. The reading results showed a bit more difference by scaling method in the θ -metric, though results were quite similar in the scale score metric. Lastly, the

[15]Scale scores are defined as expected scores on the full ECLS item pool and are, in essence, predicted observables.

within-round correlations between scores across the three scaling approaches were near 1, regardless of whether one considered the thetas or scale scores.

In summary, given the similarities of the characteristics of the aggregate score distributions and the extremely high correlations among test-taker scores, analyses carried out using the scores from any of the three of the approaches should produce highly similar findings and these should also be quite similar to those based on the scores from the earlier ECLS-K:2011 releases. Difference, if they are present, would be most likely to occur for reading analyses carried out in the θ scores.

As noted earlier, the methods employing concurrent estimation approaches (fully constrained and partially constrained) were viewed as preferable. These approaches provided a more parsimonious summarization of the regularities in the data, were better supported by the data structures (i.e., item sets administered over multiple grades and, therefore, more and better data available for item parameter estimation), and are consistent with the intended interpretation of a longitudinal scale. Moreover, there are practical advantages in the constrained approaches for the program and to potential secondary data analysts in terms of having far fewer instances of items that have multiple sets of item parameters. As part of the implementation of each of the methods, plots similar to figures 3-5 and 3-6 comparing non-parametric estimates of ICCs to the fitted logistic ICCs were produced and examined as part of the routine quality assurance and evaluation of model fit. Review of item-level model fit results for the fully constrained solution revealed in each of the subject areas significant model misfit across rounds for a number of items. Since misfit was less significant in the results using the partially constrained approach, it was selected to develop the final scale and scores. However, as part of this reestimation of the IRT scaling models, two items were removed from the final scaling and scoring due to model misfit across rounds, one from the math assessment and one from the science assessment.

3.5.2.3 Chain-Linking Methodology

A chain-linking approach was used to place item parameters and ability (theta) estimates from the kindergarten through second-grade and third- through fifth-grade scalings produced by the partially constrained solution on a common scale. TBLT is an ETS-proprietary computer program that implements the Stocking and Lord (1983) procedure. It 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

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

where a_R is the discrimination parameter for the reference scale, a_N is the discrimination parameter for the new scale, b_R is the difficulty parameter for the reference scale, b_N is the difficulty parameter for the new scale, c_R is the guessing parameter for the reference scale, and c_N is the guessing parameter for the new scale.

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

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

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.

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 fifth grade (fall kindergarten, spring kindergarten, fall first grade, spring first grade, fall second grade, spring second grade, spring third grade, spring fourth grade, and spring fifth grade) will have nine ability estimates and the associated scores for each round.

One caveat to note is the selection of item parameter estimates for the linking items was dependent upon data collection round in calculating the scale scores. As noted in section 3.5.2, for the linking items, two sets of item parameters were developed: one set for the kindergarten through second-grade calibration, and one set for the third- through fifth-grade calibration, the latter set then linked to the kindergarten through second-grade calibration scale. Calculation of the scale scores is performed by round of data collection, thus, the item parameters associated with the round of data collection were used. For example, the item "allcans6" was administered (on forms of varying difficulty) and calibrated in both the kindergarten through second-grade and third- through fifth-grade calibrations. In calculating the probability of a correct response for this particular item, the item parameters from the kindergarten through second-grade calibration were used for children with theta estimates in any of the rounds one through six (fall and spring kindergarten, fall and spring first grade, and fall and spring second grade); the item parameters from the third- through fifth-grade calibration, linked to the kindergarten through second-grade scale, were used for children with theta estimates in any of the rounds seven through nine (spring third grade, spring fourth grade, and spring fifth grade). The total number of items used in calculating the scale score remained the same across all rounds.

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

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 Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (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 ECLS-K:2011, from the initial item pool development to the finalization of the assessment forms used in the national third-, fourth-, and fifth-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 third-grade through fifth-grade assessments, prior rounds of the ECLS-K:2011 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 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 third-grade through fifth-grade assessments. Items from previous rounds of the ECLS-K:2011 were required for linking in order to continue longitudinal measurement in the study. Items from the ECLS-K allow comparisons to be made between two cohorts of kindergarten students who entered school more than a decade apart. It was also necessary to develop some new items because the existing ECLS-K:2011 and the ECLS-K 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.

4.1.1 Field-test Design

In the spring of 2013, a field test was conducted to test the assessment items being considered for inclusion in the direct child assessments for the third-, fourth-, and fifth-grade collections of the national study. This field test served as a vehicle for estimating the psychometric properties of items in the assessment battery item pool and producing psychometrically sound and valid direct cognitive assessment

instruments. The primary goal of the field test was to collect data (specifically, item statistics) to inform the development of the third-, fourth- and fifth-grade assessments for reading, mathematics, and science.

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

Data collected during the field test 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 assessments, 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.

Items field tested in the spring of 2013 were drawn from several sources: the ECLS-K:2011 second-grade assessment,[1] the ECLS-K third-grade assessment,[2] the ECLS-K fifth-grade assessment,[3] and the ECLS-K eighth-grade assessment,[4] in addition to items newly developed for the ECLS-K:2011 to measure concepts not included in the earlier studies. Since the majority of items had been used before, either in the ECLS-K:2011 or the ECLS-K, there were few concerns about their quality. However, these items were still field tested in order to calibrate the newly developed items on the same scale as the items from prior assessments. Attention was paid to the quality of the items newly developed for the ECLS-K:2011 and how the items that were previously developed performed.

In order to measure each child's ability accurately in the national assessment, it was important that each child receive a set of test items that was 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 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 third-grade national assessments

- ECLS-K:2011 2013 field test, third-graders (approximately 500 cases);
- ECLS-K:2011 2013 field test, fourth-graders (approximately 1,000 cases);

[1] Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2012 and spring 2013.
[2] Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.
[3] Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2004.
[4] Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2007.

- ECLS-K:2011 2013 field test, fifth-graders (approximately 980 cases);
- ECLS-K:2011 2013 field test, sixth-graders (approximately 480 cases);
- ECLS-K:2011 fall second-grade national data collection (approximately 4,740 cases);
- ECLS-K:2011 spring second-grade national data collection (approximately 13,850 cases);
- ECLS-K spring third-grade national data collection (approximately 14,280 cases);
- ECLS-K spring fifth-grade national data collection (approximately 11,270 cases); and
- ECLS-K spring eighth-grade national data collection (approximately 9,230 cases).

For the fourth-grade national assessment design, data from the ECLS-K:2011 third-grade national assessment were also used to inform the design. Therefore, for the fourth-grade national assessments, the following data were pooled for the analysis:

- ECLS-K:2011 2013 field test, third-graders (approximately 500 cases);
- ECLS-K:2011 2013 field test, fourth-graders (approximately 1,000 cases);
- ECLS-K:2011 2013 field test, fifth-graders (approximately 980 cases);
- ECLS-K:2011 2013 field test, sixth-graders (approximately 480 cases);
- ECLS-K:2011 fall second-grade national data collection (approximately 4,740 cases);
- ECLS-K:2011 spring second-grade national data collection (approximately 13,850 cases);
- ECLS-K:2011 spring third-grade national data collection[5] (approximately 9,800 cases);
- ECLS-K spring third-grade national data collection (approximately 14,280 cases);
- ECLS-K spring fifth-grade national data collection (approximately 11,270 cases); and
- ECLS-K spring eighth-grade national data collection (approximately 9,230 cases).

For the fifth-grade national assessment design, data from the ECLS-K:2011 fourth-grade national assessment were also used to inform the design. Therefore, for the fifth-grade national assessments, the following data were pooled for the analysis:

- ECLS-K:2011 2013 field test, third-graders (approximately 500 cases);

[5] The spring third-grade data administration was in progress during the design phase of the fourth-grade national assessments; therefore, only a portion of the spring third-grade data was available for the analysis.

- ECLS-K:2011 2013 field test, fourth-graders (approximately 1,000 cases);
- ECLS-K:2011 2013 field test, fifth-graders (approximately 980 cases);
- ECLS-K:2011 2013 field test, sixth-graders (approximately 480 cases);
- ECLS-K:2011 spring second-grade national data collection (approximately 13,850 cases);
- ECLS-K:2011 spring third-grade national data collection[6] (approximately 9,800 cases);
- ECLS-K:2011 spring fourth-grade national data collection[7] (approximately 9,300 cases);
- ECLS-K spring third-grade national data collection (approximately 14,280 cases);
- ECLS-K spring fifth-grade national data collection (approximately 11,270 cases); and
- ECLS-K spring eighth-grade national data collection (approximately 9,230 cases).

Pooling of the data for IRT calibrations was done for two primary reasons. First, for analyses of data, 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 served to stabilize parameter estimates that would lack precision if the data from only the field test were used to evaluate the assessment items.

The pool of items available for assembly of the national test forms was not limited to the items in the 2013 field test. 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 2013 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 collections.

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 deviations of the ability levels for

[6] The same data from third grade used in the design of the fourth-grade assessment were used in the fifth-grade assessment design. This was unintentional. The intent was to use the complete third-grade dataset; however, the large subset of third-grade data, in concert with the large subset of fourth-grade data, resulted in adequate data for designing the fifth-grade assessment forms.

[7] The spring fourth-grade data administration was in progress during the design phase of the fifth-grade national assessments; therefore, only a portion of the spring fourth-grade data was available for the analysis.

each of the samples were calculated based on data from the pooled sample. Therefore, an estimated ability range for the target administrations (e.g., spring third, fourth, or fifth grade) can be determined.

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 were used in prior and 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 third-, fourth-, and fifth-grade design analyses using the pooled data showed differences in the estimated mean ability levels between the ECLS-K:2011 2013 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 population in the interval between the administration of the ECLS-K national data collections and the ECLS-K:2011 2013 field test 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 2013 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 was 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-level 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 in each of the domains.

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 or she 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 or she received in the middle form might have been too difficult. 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 the child was tired and not paying close attention to questions he or 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 the child received in the middle form might have been relatively easy for him or her.

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 third-, fourth-, and fifth-grade assessments, items with low *r*-biserials, poor fit, or low discrimination were avoided.

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 third-grade means and standard deviations of reading ability level (theta), all calibrated on the same scale, for the different samples in the pooled analysis described earlier in this chapter. This information was used to calculate the full range of ability levels (and, therefore, item difficulties) that needed to be covered by the third-grade reading assessment in the ECLS-K:2011. Table 4-2 provides the same information for the fourth-grade reading assessment, and table 4-3 for the fifth-grade reading assessment. Note that the values in table 4-1 are *not* comparable to those in tables 4-2 and 4-3; 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 4-4, 4-5, and 4-6 show the estimated ability ranges for the overall assessment as well as for the low-, middle-, and high-ability-level groups for the spring of third, fourth, and fifth grades. Tables 4-4, 4-5, and 4-6 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. Not all items in the ability ranges in tables 4-4, 4-5, and 4-6 are directly comparable to each other 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 third grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring third grade – ECLS-K:2011 field test	0.10	0.91
Spring third grade – ECLS-K national data collection	0.07	0.86

NOTE: The values in this table are not comparable to the values in tables 4-2 and 4-3 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), spring 2013 field test, and Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2002.

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

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring fourth grade – ECLS-K:2011 field test	0.42	0.82
Spring fourth grade – ECLS-K national data collection (estimated)[1]	0.36	0.88

[1] Without data from the ECLS-K national data collection for fourth grade, an estimate of the spring fourth-grade ECLS-K national mean was calculated as the average of the spring third- and spring fifth-grade ECLS-K national means. The standard deviation estimate of the spring fourth-grade ECLS-K national data collection was calculated as the average of the spring third- and spring fifth-grade ECLS-K national standard deviations.
NOTE: The values in this table are not comparable to the values in tables 4-1 and 4-3 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), spring 2013 field test, and Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2002 and spring 2004.

Table 4-3. Means and standard deviations of estimated reading ability level (theta) for children in fifth grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring fifth grade – ECLS-K:2011 field test	0.51	0.74
Spring fifth grade – ECLS-K national data collection	0.53	0.82

NOTE: The values in this table are not comparable to the values in tables 4-1 and 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), spring 2013 field test, and Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2004.

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

Item	Spring third grade: 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	-1.65 to +0.10	-0.79 to +1.01	+0.07 to +1.92
Number of items with difficulties in estimated peak ability range:			
Routing	8	12	10
Low form	15	15	6
Middle form	7	13	9
High form	5	10	10

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in tables 4-5 and 4-6 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), spring 2014.

Table 4-5. Peak difficulty ranges for the national fourth-grade reading assessment, routing plus second stage: ECLS-K:2011

Item	Spring fourth grade: 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	-1.40 to +0.42	-0.52 to +1.25	+0.36 to +2.12
Number of items with difficulties in estimated peak ability range:			
Routing	8	11	9
Low form	12	8	1
Middle form	7	16	11
High form	0	7	12

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in tables 4-4 and 4-6 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), spring 2015.

Table 4-6. Peak difficulty ranges for the national fifth-grade reading assessment, routing plus second stage: ECLS-K:2011

Item	Spring fifth grade: Estimated low-level abilities (-2*SD* to mean)	Spring fifth grade: Estimated mid-level abilities (-1*SD* to +1*SD*)	Spring fifth grade: Estimated high-level abilities (mean to +2*SD*)
Estimated ability range	-1.11 to +0.53	-0.29 to +1.36	+0.51 to +2.18
Number of items with difficulties in estimated peak ability range:			
Routing	7	6	4
Low form	11	6	0
Middle form	5	14	12
High form	2	7	10

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in tables 4-4 and 4-5 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), spring 2016.

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 4-7 provides information about how the final reading assessments developed for the national third-, fourth-, and fifth-grade rounds of data collection compare to the framework specifications in terms of the distribution of items by content category. The table indicates the targeted percentage within

each content category, as well as the actual percentage and number of items selected for the national administrations within each content category.

Table 4-7. Framework targets and items by content area for the national third-grade, fourth-grade, and fifth-grade reading assessments: ECLS-K:2011

	Third grade			Fourth grade			Fifth 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**	**59**	**100**	**100**	**55**	**100**	**100**	**49**	**100**
Basic reading skills/ vocabulary	20	10	17	20	15	27	15	12	24
Comprehension									
Locate/recall	30	18	31	30	15	27	25	10	20
Integrate/ interpret	35	23	39	35	21	38	40	24	49
Critique/ evaluate	15	8	14	15	4	7	20	3	6

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), spring 2014, spring 2015, and spring 2016.

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 4-7, even with this maximization, the percentage of items in the **critique/evaluate** category fell short of the targets, especially in the fourth- and fifth-grade assessments. The available item pools did not include enough items in this category that performed well at the expected grade levels. The percentages of items in the **locate/recall**, **vocabulary**, and **integrate/interpret** categories fell at or close to the targeted percentages at each grade level, with the percentage of **vocabulary** items in fourth and fifth grades higher than targeted due to the inclusion of items in passage sets.

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, the reading assessments consisted of sets of reading items that 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 third-

grade reading assessment, 10 item sets were defined, all with a reading passage. In the fourth-grade reading assessment, 8 item sets were defined with a reading passage. And in fifth grade, 8 item sets were defined to be associated with a reading passage. The number of dichotomous and polytomous items, and for the latter, the number of score points for the third-, fourth-, and fifth-grade reading assessments are summarized in table 4-8.

Table 4-8. Dichotomous and polytomous items and score values for the national third-grade, fourth-grade, and fifth-grade reading assessments: ECLS-K:2011

Number of items	Third grade	Fourth grade	Fifth grade
Total	**24**	**27**	**21**
Dichotomous	14	19	13
Polytomous			
2-point	0	0	0
3-point	0	0	0
4-point	7	5	4
5-point	2	2	2
6-point	0	1	0
7-point	1	0	2

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014, spring 2015, and spring 2016.

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 third- and fourth-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 4-9 provides the third-grade means and standard deviations of the estimated mathematics ability level (theta), all calibrated on the same scale, for the different samples in the pooled

analysis described earlier in this chapter. This information was used to calculate the full range of ability levels (and, therefore, item difficulties) that needed to be covered by the third-grade mathematics assessment in the ECLS-K:2011. Table 4-10 provides the same information for the fourth-grade mathematics assessment, and table 4-11 provides the same information for the fifth-grade mathematics assessment. The values in tables 4-9, 4-10, and 4-11 are *not* comparable to each other; 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 4-12, 4-13, and 4-14 show the estimated ability ranges for the entire assessment as well as for the low-, middle-, and high-ability-level groups for the spring of third, fourth, and fifth grades, respectively. Tables 4-12, 4-13, and 4-14 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 4-12, 4-13, and 4-14 are not directly comparable to one another since the thetas for the third-, fourth-, and fifth-grade assessments were estimated using the different sets of pooled data described above.

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

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring third grade – ECLS-K:2011 field test	0.01	0.67
Spring third grade – ECLS-K national data collection	-0.17	0.78

NOTE: The values in this table are not comparable to the values in tables 4-10 and 4-11 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) spring 2013 field test, and Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

Table 4-10. Means and standard deviations of estimated mathematics ability level (theta) for children in fourth grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring fourth grade – ECLS-K:2011 field test	0.56	0.69
Spring fourth grade – ECLS-K national data collection (estimated)[1]	0.17	0.77

[1] Without data from the ECLS-K national data collection for fourth grade, an estimate of the spring fourth-grade ECLS-K national mean was calculated as the average of the spring third- and spring fifth-grade ECLS-K national means. The standard deviation estimate of the spring fourth-grade ECLS-K national data collection was calculated as the average of the spring third- and spring fifth-grade ECLS-K national standard deviations.
NOTE: The values in this table are not comparable to the values in tables 4-9 and 4-11 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) spring 2013 field test, and Kindergarten Class of 1998–99 (ECLS-K), spring 2002 and spring 2004.

Table 4-11. Means and standard deviations of estimated mathematics ability level (theta) for children in fifth grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring fifth grade – ECLS-K:2011 field test	0.66	0.72
Spring fifth grade – ECLS-K national data collection	0.42	0.81

NOTE: The values in this table are not comparable to the values in tables 4-9 and 4-10 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) spring 2013 field test, and Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

Table 4-12. Peak difficulty ranges for the national third-grade mathematics assessment, routing plus second stage: ECLS-K:2011

Item	Spring third grade 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	-1.73 to +0.01	-0.95 to +0.67	-0.17 to +1.34
Number of items with difficulties in estimated peak ability range:			
Routing	6	5	5
Low form	19	23	7
Middle form	8	16	13
High form	2	7	13

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in tables 4-13 and 4-14 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), spring 2014.

Table 4-13. Peak difficulty ranges for the national fourth-grade mathematics assessment, routing plus second stage: ECLS-K:2011

Item	Spring fourth grade 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	-1.40 to +0.42	-0.52 to +1.25	+0.36 to +2.12
Number of items with difficulties in estimated peak ability range:			
Routing	8	9	7
Low form	14	12	4
Middle form	10	14	9
High form	2	9	12

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in tables 4-12 and 4-14 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), spring 2015.

Table 4-14. Peak difficulty ranges for the national fifth-grade mathematics assessment, routing plus second stage: ECLS-K:2011

Item	Spring fifth grade 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	-1.20 to +0.66	-0.39 to +1.38	+0.42 to +2.10
Number of items with difficulties in estimated peak ability range:			
Routing	10	12	9
Low form	11	8	3
Middle form	9	15	10
High form	2	7	11

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in tables 4-12 and 4-13 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), spring 2016.

As discussed in chapter 2, adherence to the framework category targets in mathematics was required during assessment design. Table 4-15 provides information about how the final mathematics assessments developed for the national third-, fourth-, and fifth-grade rounds of data collection compared with the framework specifications in terms of the distribution of items by content category.

The actual percentages of items matched or were very close to the targeted percentages in all categories at all grade levels. Although the item pool included an extensive number of items of differing item types, content categories, and difficulties, not all of the items had desirable psychometric properties. Thus, some assessments produced shortfalls in actual percentages by category. 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 4-15. Framework targets and items by content area for the national third-grade, fourth-grade, and fifth-grade mathematics assessments: ECLS-K:2011

Content area	Third grade			Fourth grade			Fifth grade		
	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**	**62**	**100**	**100**	**55**	**100**	**100**	**57**	**100**
Number properties and operations	40	21	34	40	23	42	30	17	30
Measurement	20	15	24	20	11	20	18	10	18
Geometry	15	9	15	15	8	15	18	10	18
Data analysis and probability	10	6	10	10	6	11	12	7	12
Algebra	15	11	18	15	7	13	22	13	23

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), spring 2014, spring 2015, and spring 2016.

4.4 Science

Overall, the science items administered in the 2013 field test performed well. As in the reading and mathematics domains, 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 third- and fourth-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 4-16 provides the third-grade means and standard deviations of the estimated science ability level (theta), all calibrated on the same scale, for the different samples in the pooled analysis described earlier in this chapter. This information was used to calculate the full range of ability levels (and, therefore, item difficulties) that needed to be covered by the third-grade science assessment in the ECLS-K:2011. Table 4-17 provides the same information for the fourth-grade science assessment, and table 4-18 provides the same information for the fifth-grade science assessment. The values in tables 4-16, 4-17, and 4-18 are *not* comparable to each other; 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 4-19, 4-20, and 4-21 show the estimated ability ranges for the overall assessment as well as for the low-, middle-, and high-ability-level groups for the spring of third, fourth, and fifth grade, respectively. Tables 4-19, 4-20, and 4-21 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 4-19, 4-20, and 4-21 are not directly comparable to one another since the thetas for the third-, fourth-, and fifth-grade assessments were estimated using the different sets of pooled data described above.

Table 4-16. Means and standard deviations of estimated science ability level (theta) for children in third grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring third grade – ECLS-K:2011 field test	-0.12	0.79
Spring third grade – ECLS-K national data collection	-0.22	0.87

NOTE: The values in this table are not comparable to the values in tables 4-17 and 4-18 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) spring 2013 field test, and Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

Table 4-17. Means and standard deviations of estimated science ability level (theta) for children in fourth grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring fourth grade – ECLS-K:2011 field test	0.41	0.75
Spring fourth grade – ECLS-K national data collection (estimated)[1]	0.07	0.75

[1] Without data from the ECLS-K national data collection for fourth grade, an estimate of the spring fourth-grade ECLS-K national mean was calculated as the average of the spring third- and spring fifth-grade ECLS-K national means. The standard deviation estimate of the spring fourth-grade ECLS-K national data collection was calculated as the average of the spring third- and spring fifth-grade ECLS-K national standard deviations.
NOTE: The values in this table are not comparable to the values in tables 4-16 and 4-18 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) spring 2013 field test, and Kindergarten Class of 1998–99 (ECLS-K), spring 2002 and spring 2004.

Table 4-18. Means and standard deviations of estimated science ability level (theta) for children in fifth grade

Sample	Mean ability level (theta)	Standard deviation of ability level (theta)
Spring fifth grade – ECLS-K:2011 field test	0.48	0.80
Spring fifth grade – ECLS-K national data collection	0.30	0.82

NOTE: The values in this table are not comparable to the values in tables 4-16 and 4-17 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) spring 2013 field test, and Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

Table 4-19. Peak difficulty ranges for the national third-grade science assessment, routing plus second stage: ECLS-K:2011

	Spring third 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 ability range	-1.96 to -0.12	-1.09 to +0.67	-0.22 to +1.53
Number of items with difficulties in estimated peak ability range:			
Routing	6	8	8
Low form	8	10	6
Middle form	6	12	11
High form	2	7	12

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in tables 4-20 and 4-21 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), spring 2014.

Table 4-20. Peak difficulty ranges for the national fourth-grade science assessment, routing plus second stage: ECLS-K:2011

Item	Spring fourth grade 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 +0.41	-0.68 to +1.16	+0.07 to +1.91
Number of items with difficulties in estimated peak ability range:			
Routing	7	8	8
Low form	9	5	2
Middle form	7	11	8
High form	3	7	9

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in tables 4-19 and 4-21 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), spring 2015.

Table 4-21. Peak difficulty ranges for the national fifth-grade science assessment, routing plus second stage: ECLS-K:2011

Item	Spring fifth grade 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	-1.35 to +0.48	-0.52 to +1.29	+0.30 to +2.09
Number of items with difficulties in estimated peak ability range:			
Routing	7	8	9
Low form	10	4	1
Middle form	6	10	9
High form	2	7	10

NOTE: *SD* = standard deviation. The values in this table are not comparable to the values in tables 4-19 and 4-20 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), spring 2016.

Table 4-22 provides information about how the final science assessments developed for the national third-, fourth-, and fifth-grade rounds of data collection compared to the framework specifications in terms of the distribution of items by content category.

The actual percentages of items are within an item or two of the targets in all categories. 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 4-22. Framework targets and items by content area for the national third-grade, fourth-grade, and fifth-grade science assessments: ECLS-K:2011

	Third grade			Fourth grade			Fifth 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**	**47**	**100**	**100**	**40**	**100**	**100**	**40**	**100**
Scientific inquiry	25	12	26	25	10	25	25	10	25
Physical science	25	9	19	25	10	25	25	10	25
Life science	25	13	28	25	10	25	25	10	25
Earth and space science	25	13	28	25	10	25	25	10	25

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014, spring 2015, and spring 2016.

4.5 Cut Scores Used for Routing

As noted earlier, the third-, fourth-, and fifth-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 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 third grade included simulations on data from two samples: (1) the spring of third grade from the ECLS-K:2011 field test (interpolated from the fall of third-grade and the fall of fourth-grade data) and (2) the spring of third grade from the ECLS-K national data collection. The cut score simulations for fourth grade also included data from two samples: (1) the spring of fourth grade from the ECLS-K:2011 field test (interpolated from the fall of fourth-grade and the fall of fifth-grade data) and (2) the spring of fourth grade from the ECLS-K national data collection (interpolated from the spring of third-grade and the spring of fifth-grade data). The two samples used in the simulations for fifth grade were (1) the spring of fifth grade from the ECLS-K:2011 field test (interpolated from the fall of fifth-grade and the fall of sixth-grade data) and (2) the spring of fifth grade from the ECLS-K national data collection.

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, two simulations were performed for reading, one for each routing form 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 third-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 8 or lower (including 0). Children who scored between 9 and 14 items (inclusive) correct on the router would proceed with the middle form, while those with scores of 15 or higher would proceed to the high form. One additional routing rule was included, which routed children who incorrectly responded to the first four items immediately to the low form. The first four items required children to read a sentence and fill in the blank. If all four of the sentence-reading items were incorrect, the child was not administered any of the subsequent router items, all of which were associated with passages.

For the fourth-grade reading assessment, the analysis of optimal cut scores indicated that children with a router score of 8 or lower should be directed to the low form, while children with router scores of 9 to 14 and 15 or more should be directed to the middle and high forms, respectively.

Lastly, for the fifth-grade reading assessment, analysis indicated that children with a router score of 4 or lower should be directed to the low form, children with a router score between 5 and 8 should be directed to the middle form, and children with a router score of 9 or higher should be directed to the high form.

4.5.2 Mathematics

For the third-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 11 items (inclusive) correct on the router would proceed with the middle form, while those with scores of 12 or higher would proceed to the high form.

For the fourth-grade mathematics assessment, the analysis of optimal cut scores indicated that children with a router score of 7 or lower should be directed to the low form, while children with router scores of 8 to 12 and 13 to 17 should be directed to the middle and high forms, respectively.

Lastly, for the fifth-grade mathematics assessment, analysis indicated children with a router score of 7 or lower should be directed to the low form, children with a router score between 8 and 12 should be directed to the middle form, and children with a router score of 13 or higher should be directed to the high form.

4.5.3 Science

For the third-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 5 or lower. Children who scored between 6 and 9 items (inclusive) 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 fourth-grade science assessment, the analysis of optimal cut scores indicated that children with a router score of 6 or lower should be directed to the low form, while children with router scores of 7 to 10 and 11 or more should be directed to the middle and high forms, respectively.

For the fifth-grade science assessment, analysis indicated that children with a router score of 6 or lower should be directed to the low form, children with a router score between 7 and 10 should be directed to the middle form, and children with a router score of 11 or higher should be directed to the high form.

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 answered the sentence-reading items incorrectly most likely would not be able to answer the passage-associated items correctly. 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.

Discontinue rules in the mathematics and science assessments 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 administrations. Based on these estimates and guided by professional judgments, 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 third-grade reading assessment, one discontinue rule was implemented:

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

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

1. On the routing form, if the first two read-aloud words were answered incorrectly, the router was ended and the child was directed to a second-stage form.
2. On the routing form, if four of the first six read-aloud words 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 three of the four items associated with the first passage were answered incorrectly, the reading assessment was discontinued.
4. On the second-stage high form, if two of the four items associated with the second passage were answered incorrectly, the reading assessment was discontinued.

There were no discontinue rules implemented in the fifth-grade reading assessment.

4.6.2 Mathematics

In the third-grade mathematics assessment, one discontinue rule were implemented:

1. On the routing form, if at least 4 of the first 12 items were answered incorrectly, the router was ended and the child was directed to a second-stage form.

In the fourth-grade mathematics assessment, four discontinue rules were implemented:

1. On the routing form, if three of four middle-difficulty items were answered incorrectly, the router was ended and the child was directed to a second-stage form.
2. On the second-stage low form, if two of three middle-difficulty items were answered incorrectly, the mathematics assessment was discontinued.
3. On the second-stage middle form, if three of four middle-difficulty items were answered incorrectly, the mathematics assessment was discontinued.
4. On the second-stage high form, if three of four middle-difficulty items were answered incorrectly, the mathematics assessment was discontinued.

There were no discontinue rules implemented in the fifth-grade mathematics assessment.

4.6.3 Science

In the third-grade science assessment, three discontinue rules were implemented:

1. On the second-stage low form, if at least 7 of the first 10 items were answered incorrectly, the science assessment was discontinued.

2. On the second-stage middle form, if at least 8 of the first 12 items were answered incorrectly, the science assessment was discontinued.

3. On the second-stage high form, if at least 10 of the first 13 items were answered incorrectly, the science assessment was discontinued.

There were no discontinue rules implemented in the fourth- and fifth-grade science assessments.

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, fall 2012 and spring 2013 second-grade, spring 2014 third-grade, spring 2015 fourth-grade, and spring 2016 fifth-grade rounds of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011). Although the focus of this report is predominantly on the psychometric results for the third-, fourth-, and fifth-grade data collections, certain information (e.g., estimates in the tables) from the kindergarten, first-grade, and second-grade data collection rounds are also included in this chapter to provide the reader with the complete longitudinal analysis results. For additional information on the kindergarten, first-grade, and second-grade data collection rounds, see the *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a) and the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First-Grade and Second-Grade Psychometric Report.* (NCES 2018-183) (Najarian et al. 2018b). 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 5.5 focus on the reading, mathematics, and science assessments, respectively, followed by an evaluation of the longitudinal scale presented in section 5.6. The chapter concludes with a discussion of the selection and use of the scores in section 5.7, with section 5.7.1 focusing on choosing the best scores for certain types of analyses and section 5.7.2 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. In the kindergarten rounds of the

[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 (see Diamond 2013). Two measures of executive function were administered in fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: The *Dimensional Change Card Sort (DCCS),* which measures cognitive flexibility, and the Numbers Reversed subtask from the *Woodcock-Johnson III Tests of Cognitive Abilities*, which measures working memory. In fourth grade, a third measure of executive function was added. The *Flanker* task, which measures inhibitory control in the context of selective visual attention, was administered in fourth grade and fifth grade. Executive function measures are discussed in more detail in chapter 6.

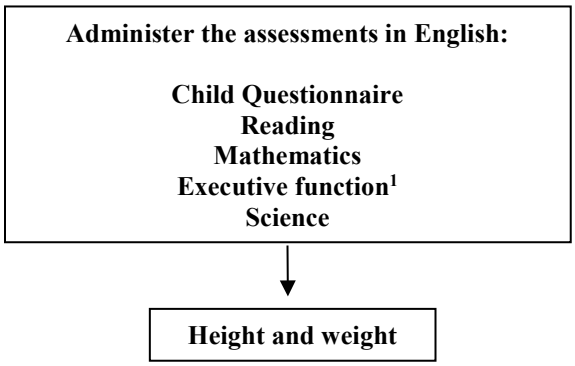
assessment, all children were administered an English-language screener, regardless of home language. For children whose primary home language was English, the screener served as a warm-up or practice for the rest of the assessment. Children's performance on the language screener determined the routing through the assessment and the language of administration. Children who achieved at least a minimum score on the screener were administered the entire assessment in English. Children whose home language was not English and did not achieve the minimum score or higher received certain components of the assessment in Spanish if Spanish was their home language or were routed out of the assessment and had only their height and weight measured if Spanish was not their home language. In first grade, depending on the path in kindergarten, children were re-administered the language screener. Again, performance on the screener determined the routing and language of administration of the assessment. By the spring of first grade, nearly all children had achieved at least the minimum score on the language screener, so it was not necessary to include the screener or the assessments in Spanish in the later rounds of data collection. For more information on the administration of the kindergarten, first-, and second-grade assessments, see 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. 2015a); the *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) (Tourangeau et al. 2015b); and the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES 2017-285) (Tourangeau et al. 2017).

5.1.1 Third-Grade Round

In the third-grade data collection, children were assessed in reading, mathematics, and science in the spring only. All children received the assessments designed for the third-grade collections, regardless of their actual grade level. Starting in third grade, students completed a child questionnaire. Students' executive function skills were assessed with the same measures fielded in kindergarten and first grade, the *DCCS*[2] and Numbers Reversed. Finally, children's height and weight were measured. Exhibit 5-1 illustrates the administration of the third-grade assessment for all children, with the domains listed in order of administration.

[2] Though the *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.

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

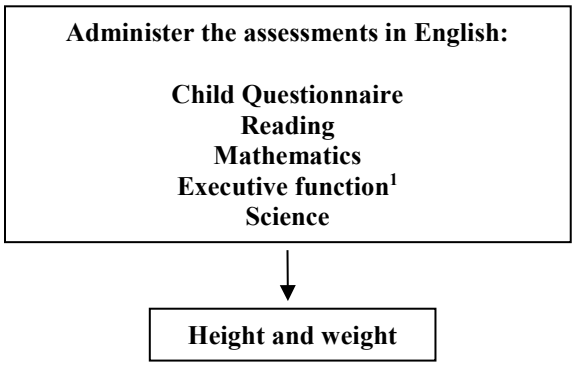


[1] Two measures of executive function were administered in third grade. The *DCCS* was administered first, followed by Numbers Reversed.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014.

5.1.2 Fourth-Grade Round

The fourth-grade data collection followed the design of the third-grade data collection, with children assessed in reading, mathematics, and science in the spring only. All children received the assessments designed for the fourth-grade collections, regardless of their actual grade level. A child questionnaire was administered, with new content along with some content that was repeated from third grade. Students' executive function skills were assessed with the same measures fielded in prior rounds, the *DCCS* and Numbers Reversed. In addition, a new measure of executive function, the *Flanker*, was administered for the first time. Finally, children's height and weight were measured. Exhibit 5-2 illustrates the administration of the fourth-grade assessment for all children, with the domains listed in order of administration.

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

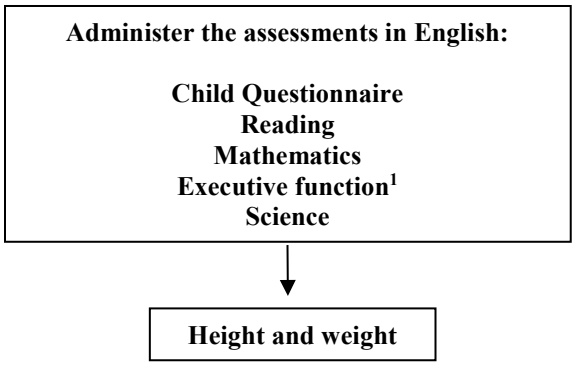


[1] Three measures of executive function were administered in fourth grade. The *DCCS* was administered first, followed by Numbers Reversed. The *Flanker* was administered for the first time in fourth grade, and it was administered after Numbers Reversed.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2015.

5.1.3 Fifth-Grade Round

The fifth-grade data collection followed the design of the third- and fourth-grade data collections, with children assessed in reading, mathematics, and science in the spring only. All children received the assessments designed for the fifth-grade collections, regardless of their actual grade level. Students completed a child questionnaire, which had varying content across the third-, fourth-, and fifth-grade rounds of data collection. Students' executive function skills were assessed with the same three measures fielded in fourth grade, the *DCCS,* Numbers Reversed, and *Flanker*. Finally, children's height and weight were measured. Exhibit 5-3 illustrates the administration of the fifth-grade assessment 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 fifth-grade year



[1] Three measures of executive function were administered in fifth grade. The *DCCS* was administered first, followed by Numbers Reversed and then the *Flanker*.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2016.

5.2 Scoring the National Assessment

This section presents information about the assessment scores developed for the kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-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. Scores comparable across all rounds 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.

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.

Component analyses are a typical approach for looking at unidimensionality. There is no definitive rule of thumb, but generally the larger the eigenvalue for the first factor in comparison to the second factor, the more evidence of unidimensionality. The goal is to see a dominant first factor that correlates with the latent trait being measured such as reading, mathematics, and science. Components analyses can sometimes have a strong second factor that is related to item difficulty. This, however, is not inconsistent with a unidimensional dominant trait, because item difficulty is generally not construct relevant and does not suggest the items are loading on multiple constructs.

As noted in chapter 3, the ideal definition of unidimensionality is when the ratios of the first component to the second component and of the first component to the third component are both at least 3:1. Dividing the percentage of the second component by the percentage of the first component (or dividing the percentage of the third component by the percentage of the first component) provides a metric for comparison. For example, if the first component is equal to 16 and the second component is equal to 4, dividing the two equals 0.25, and thus satisfies the ideal threshold for unidimensionality since the result (4/16 = 0.25) is less than 0.33 (or 1/3).

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 (33 percent to 60 percent of the first component) (tables 5-1 and 5-2). In second grade, the reading assessment showed a large single component with a second component higher than expected (40 percent of the first component) and a third component at expected levels for unidimensionality (25 percent of the first component) (table 5-3). In third, fourth, and fifth grades, a large single component was followed by a comparably sized second component and a relatively smaller third component (second components 77 percent to 94 percent of the first component, and third components 43 percent to 69 percent of the first component) (tables 5-4, 5-5, and 5-6).

For mathematics, the component analyses for each grade showed a large single component but with second components that represented higher than expected percentages for unidimensionality (44

percent to 64 percent of the first component) and third components that were within expected values for the second- through fifth-grade rounds (33 percent or less of the first component) but higher than expected for the kindergarten and first-grade rounds (41 percent to 42 percent of the first component) (tables 5-1 through 5-6).

For science, the component analyses across all grades (tables 5-1 through 5-6) showed a large single component, a second component higher than expected for unidimensionality (33 percent to 58 percent of the first component), and a third component within the expected range (28 percent to 32 percent of the first component) for all rounds except spring third grade (39 percent of the first component).

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 and domain, ECLS-K:2011 fall and spring kindergarten data collections: School year 2010–11

Domain	Percent of component 1	Percent of component 2	Percent of component 3	Component 2 percent of component 1	Component 3 percent of component 1
Reading	13.62	8.17	5.14	59.99	37.74
Mathematics	10.73	6.58	4.43	61.32	41.29
Science	18.64	6.22	5.52	33.37	29.61

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 2010 and spring 2011.

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

Domain	Percent of component 1	Percent of component 2	Percent of component 3	Component 2 percent of component 1	Component 3 percent of component 1
Reading	10.42	5.36	3.44	51.44	33.01
Mathematics	10.23	6.53	4.32	63.83	42.23
Science	12.05	4.97	3.38	41.24	28.05

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 and domain, ECLS-K:2011 fall and spring second-grade data collections: School year 2012–13

Domain	Percent of component 1	Percent of component 2	Percent of component 3	Component 2 percent of component 1	Component 3 percent of component 1
Reading	19.27	7.79	4.88	40.43	25.32
Mathematics	11.26	6.64	3.74	58.97	33.21
Science	9.98	5.43	3.14	54.41	31.46

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.

Table 5-4. Component analysis percentages by component and domain, ECLS-K:2011 spring third-grade data collection: Spring 2014

Domain	Percent of component 1	Percent of component 2	Percent of component 3	Component 2 percent of component 1	Component 3 percent of component 1
Reading	3.92	3.00	1.70	76.53	43.37
Mathematics	6.82	3.53	2.10	51.76	30.79
Science	4.28	2.35	1.66	54.91	38.79

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), spring 2014.

Table 5-5. Component analysis percentages by component and domain, ECLS-K:2011 spring fourth-grade data collection: Spring 2015

Domain	Percent of component 1	Percent of component 2	Percent of component 3	Component 2 percent of component 1	Component 3 percent of component 1
Reading	3.47	3.00	2.39	86.46	68.88
Mathematics	6.75	3.30	2.23	48.89	33.04
Science	4.39	2.55	1.40	58.09	31.89

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), spring 2015.

Table 5-6. Component analysis percentages by component and domain, ECLS-K:2011 spring fifth-grade data collection: Spring 2016

Domain	Percent of component 1	Percent of component 2	Percent of component 3	Component 2 percent of component 1	Component 3 percent of component 1
Reading	3.00	2.83	1.73	94.33	57.67
Mathematics	7.56	3.32	2.06	43.92	27.25
Science	4.69	2.41	1.42	51.39	30.28

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), spring 2016.

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 do 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 (e.g., spring third grade, spring fourth grade, and spring fifth 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.) For information on the DIF analysis for items in the kindergarten, first-grade, and second-grade assessments, see the *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a) and the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b).

The DIF categories are based on differences expressed on the delta scale of item difficulty, known as Mantel-Haenszel Delta Difference (M-H D-DIF). Category A consists of items with M-H D-DIF not significantly different from zero, or less than 1.0 in absolute value. Category C consists of items with M-H D-DIF significantly greater than 1.0 and absolute value of 1.5 or greater. Category B consists of all other items. The tables in this section present information on the items identified with C-level M-H D-DIF and those with primary item discrepancy index (P-DIF). Within the table note, the item(s) presented in the table are identified. Appendix C of this report presents a table of all the assessment items, when they were fielded, and their IRT item parameters.

5.2.2.1 Third-Grade Round

As in the prior rounds, M-H D-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 5-7 summarizes the results of the M-H D-DIF and P-DIF analyses for all reading items. Both C-level M-H D-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 5-7. Reading assessment differential item functioning, ECLS-K:2011 spring third-grade data collection: Spring 2014

Result	Comparison			
	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. Item favoring White children in the White/Asian contrast: THOSEDAY (from the polytomous item set CL_KA).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014.

Table 5-8 summarizes the results of the DIF analysis for the spring third-grade round in mathematics. No items exhibited borderline C-level DIF and P-DIF.

Table 5-8. Mathematics assessment differential item functioning, ECLS-K:2011 spring third-grade data collection: Spring 2014

Result	Comparison			
	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	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), spring 2014.

DIF procedures were also used to analyze the third-grade science assessment. Table 5-9 summarizes the results in science. Two items exhibited both C-level DIF and P-DIF: one against females and one against males. Upon review, these items were retained since it was determined they did not exhibit any observable bias.

Table 5-9. Science assessment differential item functioning, ECLS-K:2011 spring third-grade data collection: Spring 2014

Result	Comparison			
	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	1	0	0	0
Number of DIF items favoring focal group	1	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. Item favoring male children in the male/female contrast: SOLARCAR. Item favoring female children in the male/female contrast: MAMMALS.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014.

5.2.2.2 Fourth-Grade Round

As in the prior rounds, the M-H D-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. Table 5-10 summarizes the results of the M-H D-DIF and P-DIF analyses for all reading items. Both C-level M-H D-DIF and P-DIF against two race/ethnicity focal groups were observed for five items. Upon review, these items were retained since bias was not indicated.

Table 5-10. Reading assessment differential item functioning, ECLS-K:2011 spring fourth-grade data collection: Spring 2015

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	0	0	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). DIF = differential item functioning. Item favoring Hispanic children in the White/Hispanic contrast: DOMESTIC. Items favoring White children in the White/Asian contrast: THOSEDAY (from the polytomous item set CL_KA), and WHYNOT, SIMPROB, and HELPUND (from the polytomous item set CL_WJ).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2015.

Table 5-11 summarizes the results of the DIF analysis for the spring fourth-grade round in mathematics. One item exhibited borderline C-level DIF and P-DIF against males, but it was retained for scoring since it was not determined to exhibit any observable bias.

Table 5-11. Mathematics assessment differential item functioning, ECLS-K:2011 spring fourth-grade data collection: Spring 2015

	Comparison			
Result	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	0	0	0
Number of DIF items favoring focal group	1	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. Item favoring female children in the male/female contrast: MYSTERY.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2015.

DIF procedures were also used to analyze the fourth-grade science assessment. Table 5-12 summarizes the results in science. None of the items exhibited both C-level DIF and P-DIF.

Table 5-12. Science assessment differential item functioning, ECLS-K:2011 spring fourth-grade data collection: Spring 2015

Result	Comparison			
	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	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), spring 2015.

5.2.2.3 Fifth-Grade Round

As in the prior rounds, the M-H D-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. Table 5-13 summarizes the results of the M-H D-DIF and P-DIF analyses for all reading items. None of the items exhibited both C-level M-H D-DIF and P-DIF.

Table 5-13. Reading assessment differential item functioning, ECLS-K:2011 spring fifth-grade data collection: Spring 2016

Result	Comparison			
	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	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), spring 2016.

Table 5-14 summarizes the results of the DIF analysis for the spring fifth-grade round in mathematics. None of the items exhibited C-level DIF and P-DIF.

Table 5-14. Mathematics assessment differential item functioning, ECLS-K:2011 spring fifth-grade data collection: Spring 2016

Result	Comparison			
	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	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), spring 2016.

DIF procedures were also used to analyze the fifth-grade science assessment. Table 5-15 summarizes the results in science. None of the items exhibited both C-level DIF and P-DIF.

Table 5-15. Science assessment differential item functioning, ECLS-K:2011 spring fifth-grade data collection: Spring 2016

Result	Comparison			
	Male/ female	White/ Black	White/ Hispanic	White/ Asian
Number of DIF items favoring reference group	0	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), spring 2016.

5.2.3 Assessment Score Reliability

Estimates of the reliability for each score (by domain by assessment year) of the overall IRT ability estimate (Lord 1980) are presented in tables 5-20, 5-25, and 5-32 later in this chapter. 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[3] 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, second grade, third grade, fourth grade, and fifth 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 assessment (e.g., the kindergarten science assessment) because there were no discontinue rules or routing into second-stage tests with different items, in practice not all children have responses for all items in these assessments. Omissions by the child or the discontinuation of the assessment (for example, if a child became too tired to continue or refused to answer) resulted in some children who began the single-stage assessments having missing data for some items. In these cases, IRT was used to estimate the child's probability of a correct response when no response information was available. IRT uses the pattern of right, wrong, and

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

omitted responses to the items actually administered in an assessment and the difficulty, discriminating ability,[4] and "guess-ability" of each item to estimate each child's ability on the same continuous scale.

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

As discussed in chapter 3, the IRT methodology used to calibrate the assessment data and produce scores was modified once all of the data for all of the rounds of administration was collected. Correlations between the scores produced from prior calibrations and those produced in the revised calibrations are extremely high, ranging from .97 to 1.0 for reading and .99 to 1.0 for mathematics and science theta scores. Correlations are between .93 and 1.0 for reading scale scores and between 0.99 and 1.00 for scale scores in both mathematics and science. Rescaling of the theta and scale scores is highly unlikely to change the substantive findings from any analyses conducted in the past using the previous scalings.

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

The theta score is an estimate of a child's ability in a particular domain (e.g., reading, mathematics, or science) based on that child's performance on the items administered. This score represents a child's latent ability and is not dependent on the difficulty of the items a child was administered. Theta scores 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.

[4] 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.

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, the methodology for the scaling of the reading, mathematics, and science direct child assessments was reexamined and, based on these findings, a partially constrained approach was selected. The partially constrained solution incorporates both concurrent calibration and chain linking to develop scores allowing for longitudinal measurement within and across grades. Two separate concurrent IRT calibrations were conducted—one for the combined dataset from the first six rounds of data collection (fall and spring kindergarten, fall and spring first grade, and fall and spring second grade) and one for the combined dataset from rounds 7 through 9 (spring third, spring fourth, and spring fifth grades). Within each of these two calibrations, separate ability distributions were estimated for each data collection round but a single common item characteristic curve was estimated for each item. The results of these separate scalings were then linked to place the third-, fourth-, and fifth-grade calibration results on the kindergarten through second-grade scale. Unlike prior data deliveries, *all* scores (including theta and the standard error of theta) have been recalibrated for each round; thus, analyses using *any* of the fifth-grade scores will be comparable only with scores from other data collection rounds provided in the kindergarten through fifth-grade data file.

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, or science) are summed to create the overall scale score. Because the computed scale scores are sums of probabilities, the scores are not integers.

The probability that a child would have gotten an item correct is dependent on the difficulty, discrimination, and guessing parameter estimates of the item, as well as the ability estimate (theta) of the child. For example, in an item set designed for both 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 fifth-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.

5.3 Reading Assessment

5.3.1 Samples and Associated Statistics for the Third-Grade Round[5]

The third-grade reading assessment consisted of 20 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 22, 16, and 17 items, respectively. Discontinue rules were employed 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. For information on discontinue rules, see chapter 4, section 4.6.

The total number of children who were administered the reading assessment and the assessment's associated statistics are shown in table 5-16. There was no evidence of a floor effect or a ceiling effect in the spring administration. Review of the classical item analysis *r*-biserials in the spring showed all items above the value of .3.

[5] For samples and associated statistics for the kindergarten, first-grade, and second-grade rounds, see the *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a) and the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b).

Table 5-16. Third-grade reading assessment sample, ECLS-K:2011 spring third-grade data collections: Spring 2014

Characteristic	Spring third grade Number	Spring third grade Percent
Total sample size	**12,897**	**100**
Number of children with responses to fewer than 10 items	31	#
Number of children with scoreable data	12,866	100
Number and percent of children with responses only for the router	1	#
Number and percent of children routed to low form	3,666	28
Number and percent of children routed to middle form	6,607	51
Number and percent of children routed to high form	2,592	20
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 + low form	166	1

Rounds to zero.
NOTE: The unweighted *n* is the 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), spring 2014.

5.3.2 Samples and Associated Statistics for the Fourth-Grade Round

The fourth-grade reading assessment consisted of 19 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 16, 18, and 15 items, respectively. Discontinue rules were employed as in prior rounds. For information on discontinue rules, see chapter 4, section 4.6.

The total number of children who were administered the reading assessment and the assessment's associated statistics are shown in table 5-17. There was no evidence of a floor effect or a ceiling effect. Review of the classical item analysis *r*-biserials showed two items with slightly lower *r*-biserials than the ideal minimum of .3: both items were very difficult for the majority of the sample.

Table 5-17. Fourth-grade reading assessment sample, ECLS-K:2011 spring fourth-grade data collection: Spring 2015

Characteristic	Spring fourth grade	
	Number	Percent
Total sample size	**12,103**	**100**
Number of children with responses to fewer than 10 items	29	#
Number of children with scoreable data	12,074	100
Number and percent of children with responses only for the router	1	#
Number and percent of children routed to low form	3,272	27
Number and percent of children routed to middle form	5,634	47
Number and percent of children routed to high form	3,167	26
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 + low form	138	1

Rounds to zero.
NOTE: The unweighted *n* is the 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), spring 2015.

5.3.3 Samples and Associated Statistics for the Fifth-Grade Round

The fifth-grade reading assessment consisted of 12 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 19, 18, and 17 items, respectively. Discontinue rules were employed as in prior rounds. For information on discontinue rules, see chapter 4, section 4.6.

The total number of children who were administered the reading assessment and the assessment's associated statistics are shown in table 5-18. There was no evidence of floor or ceiling effects in spring administration. Review of the classical item analysis *r*-biserials showed all items with *r*-biserials greater than .3.

Table 5-18. Fifth-grade reading assessment samples, ECLS-K:2011 spring fifth-grade data collection: Spring 2016

Characteristic	Spring fifth grade	
	Number	Percent
Total sample size	**11,445**	**100**
Number of children with responses to fewer than 10 items	18	#
Number of children with scoreable data	11,427	100
Number and percent of children with responses only for the router	0	0
Number and percent of children routed to low form	1,399	12
Number and percent of children routed to middle form	5,303	46
Number and percent of children routed to high form	4,725	41
Number and percent of children with a perfect score: router + high form	1	#
Number and percent of children with a chance score or below: router + low form	63	1

Rounds to zero.
NOTE: The unweighted *n* is the 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), spring 2016.

5.3.4 Score Statistics

Table 5-19 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 167 unique items administered in the kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade rounds. The IRT-based scores are calculated for all children with scoreable reading assessment data. Plots of the IRT theta and scale score distributions are provided in appendix D. Tables 5-20 and 5-21 present summary statistics by selected child characteristics for the IRT-based reading theta and scale scores in the kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade rounds.

Table 5-19. Reading assessment statistics, by IRT-based score, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

Variable	Description	*n*	Range of possible values	Weighted mean	Weighted standard deviation
X1RTHETK5	X1 READING THETA-K5	15,669	-4.0–+4.0	-1.24	0.792
X2RTHETK5	X2 READING THETA-K5	17,186	-4.0–+4.0	-0.29	0.660
X3RTHETK5	X3 READING THETA-K5	5,194	-4.0–+4.0	0.05	0.584
X4RTHETK5	X4 READING THETA-K5	15,115	-4.0–+4.0	0.55	0.493
X5RTHETK5	X5 READING THETA-K5	4,725	-4.0–+4.0	0.72	0.402
X6RTHETK5	X6 READING THETA-K5	13,837	-4.0–+4.0	0.94	0.360
X7RTHETK5	X7 READING THETA-K5	12,866	-4.0–+4.0	1.12	0.300
X8RTHETK5	X8 READING THETA-K5	12,074	-4.0–+4.0	1.29	0.295
X9RTHETK5	X9 READING THETA-K5	11,427	-4.0–+4.0	1.45	0.346
X1RSETHK5	X1 READING STD ERR OF THETA-K5	15,669	0.0–+3.0	0.30	0.083
X2RSETHK5	X2 READING STD ERR OF THETA-K5	17,186	0.0–+3.0	0.20	0.080
X3RSETHK5	X3 READING STD ERR OF THETA-K5	5,194	0.0–+3.0	0.16	0.061
X4RSETHK5	X4 READING STD ERR OF THETA-K5	15,115	0.0–+3.0	0.12	0.040
X5RSETHK5	X5 READING STD ERR OF THETA-K5	4,725	0.0–+3.0	0.11	0.023
X6RSETHK5	X6 READING STD ERR OF THETA-K5	13,837	0.0–+3.0	0.11	0.016
X7RSETHK5	X7 READING STD ERR OF THETA-K5	12,866	0.0–+3.0	0.11	0.010
X8RSETHK5	X8 READING STD ERR OF THETA-K5	12,074	0.0–+3.0	0.11	0.016
X9RSETHK5	X9 READING STD ERR OF THETA-K5	11,427	0.0–+3.0	0.13	0.030
X1RSCALK5	X1 READING IRT SCALE SCORE-K5	15,669	0.0–167.0	53.85	11.224
X2RSCALK5	X2 READING IRT SCALE SCORE-K5	17,186	0.0–167.0	68.57	14.315
X3RSCALK5	X3 READING IRT SCALE SCORE-K5	5,194	0.0–167.0	77.03	16.715
X4RSCALK5	X4 READING IRT SCALE SCORE-K5	15,115	0.0–167.0	94.47	17.812
X5RSCALK5	X5 READING IRT SCALE SCORE-K5	4,725	0.0–167.0	101.22	17.413
X6RSCALK5	X6 READING IRT SCALE SCORE-K5	13,837	0.0–167.0	111.93	16.922
X7RSCALK5	X7 READING IRT SCALE SCORE-K5	12,866	0.0–167.0	120.66	15.331
X8RSCALK5	X8 READING IRT SCALE SCORE-K5	12,074	0.0–167.0	129.31	14.513
X9RSCALK5	X9 READING IRT SCALE SCORE-K5	11,427	0.0–167.0	136.26	15.337

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the 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, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 5-20. Reading assessment theta statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

	Fall kindergarten (X1RTHETK5)			Spring kindergarten (X2RTHETK5)			Fall first grade (X3RTHETK5)			Spring first grade (X4RTHETK5)			Fall second grade (X5RTHETK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**15,669**	**-1.24**	**0.792**	**17,186**	**-0.29**	**0.660**	**5,194**	**0.05**	**0.584**	**15,115**	**0.55**	**0.493**	**4,725**	**0.72**	**0.402**
Sex															
Male	7,988	-1.29	0.809	8,752	-0.34	0.694	2,711	-0.02	0.604	7,703	0.50	0.520	2,451	0.68	0.420
Female	7,646	-1.18	0.769	8,403	-0.23	0.616	2,480	0.11	0.555	7,390	0.61	0.456	2,273	0.76	0.379
Race/ethnicity															
White, non-Hispanic	7,609	-1.09	0.748	8,076	-0.17	0.572	1,936	0.13	0.551	7,098	0.64	0.442	1,753	0.80	0.385
Black, non-Hispanic	2,114	-1.35	0.742	2,224	-0.41	0.660	538	-0.05	0.586	1,803	0.45	0.499	462	0.62	0.388
Hispanic	3,784	-1.54	0.781	4,366	-0.52	0.743	1,973	-0.11	0.599	3,987	0.40	0.544	1,857	0.59	0.414
Asian, non-Hispanic	1,174	-0.91	0.898	1,437	-0.08	0.701	402	0.31	0.593	1,302	0.69	0.467	363	0.84	0.286
Hawaiian, Other Pacific Islander, non-Hispanic	86	-1.44	0.907	114	-0.34	0.754	29	-0.17	0.864	99	0.53	0.561	23	0.66	0.402
American Indian/ Alaska Native, non-Hispanic	143	-1.54	0.752	149	-0.51	0.660	97	-0.09	0.490	129	0.47	0.445	81	0.60	0.349
Two or more races, non-Hispanic	722	-1.13	0.788	783	-0.20	0.617	211	0.07	0.631	682	0.60	0.473	181	0.77	0.428
School type (fall kindergarten)															
Public school	13,581	-1.26	0.793	13,529	-0.31	0.663	4,187	0.01	0.570	12,022	0.54	0.492	3,840	0.70	0.402
Private school	2,088	-1.00	0.744	2,060	-0.15	0.614	480	0.31	0.546	1,640	0.66	0.451	415	0.87	0.363

See notes at end of table.

Table 5-20. Reading assessment theta statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016—Continued

	Spring second grade (X6RTHETK5)			Spring third grade (X7RTHETK5)			Spring fourth grade (X8RTHETK5)			Spring fifth grade (X9RTHETK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**13,837**	**0.94**	**0.360**	**12,866**	**1.12**	**0.300**	**12,074**	**1.29**	**0.295**	**11,427**	**1.45**	**0.346**
Sex												
Male	7,038	0.90	0.384	6,569	1.09	0.316	6,150	1.27	0.316	5,839	1.44	0.364
Female	6,780	0.98	0.326	6,280	1.15	0.279	5,909	1.31	0.270	5,575	1.47	0.324
Race/ethnicity												
White, non-Hispanic	6,534	1.01	0.333	6,092	1.18	0.286	5,705	1.35	0.278	5,421	1.53	0.326
Black, non-Hispanic	1,577	0.84	0.359	1,413	1.00	0.277	1,280	1.17	0.276	1,115	1.31	0.334
Hispanic	3,739	0.83	0.379	3,527	1.02	0.303	3,369	1.20	0.297	3,252	1.35	0.340
Asian, non-Hispanic	1,194	1.03	0.307	1,111	1.19	0.265	1,045	1.38	0.283	1,002	1.56	0.328
Hawaiian, Other Pacific Islander, non-Hispanic	78	0.96	0.331	72	1.09	0.265	67	1.28	0.275	62	1.45	0.291
American Indian/ Alaska Native, non-Hispanic	114	0.85	0.385	106	1.05	0.292	105	1.24	0.293	101	1.39	0.366
Two or more races, non-Hispanic	590	1.00	0.359	535	1.18	0.299	494	1.34	0.307	465	1.51	0.362
School type (fall kindergarten)												
Public school	11,063	0.93	0.359	10,315	1.11	0.302	9,661	1.28	0.294	9,146	1.44	0.345
Private school	1,468	1.06	0.309	1,353	1.22	0.257	1,276	1.40	0.269	1,199	1.58	0.318

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimate (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the number of cases with a valid score. 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 2010, spring 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 5-21. Reading assessment scale score statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

	Fall kindergarten (X1RSCALK5)			Spring kindergarten (X2RSCALK5)			Fall first grade (X3RSCALK5)			Spring first grade (X4RSCALK5)			Fall second grade (X5RSCALK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**15,669**	**53.85**	**11.224**	**17,186**	**68.57**	**14.315**	**5,194**	**77.03**	**16.715**	**15,115**	**94.47**	**17.812**	**4,725**	**101.22**	**17.413**
Sex															
Male	7,988	53.34	11.446	8,752	67.66	14.460	2,711	75.45	16.534	7,703	92.59	18.202	2,451	99.60	17.785
Female	7,646	54.40	10.958	8,403	69.53	14.081	2,480	78.70	16.743	7,390	96.46	17.166	2,273	102.93	16.844
Race/ethnicity															
White, non-Hispanic	7,609	55.59	11.258	8,076	70.86	14.312	1,936	79.15	16.992	7,098	97.77	17.191	1,753	104.74	17.346
Black, non-Hispanic	2,114	52.21	9.807	2,224	65.85	12.956	538	74.23	15.806	1,803	90.45	17.012	462	96.59	15.665
Hispanic	3,784	50.11	9.459	4,366	64.17	12.652	1,973	72.80	14.922	3,987	88.63	17.586	1,857	95.35	16.672
Asian, non-Hispanic	1,174	59.61	15.811	1,437	74.69	17.822	402	86.23	18.475	1,302	100.41	16.918	363	106.35	13.864
Hawaiian, Other Pacific Islander, non-Hispanic	86	51.92	12.045	114	68.36	16.130	29	74.27	19.058	99	94.40	18.351	23	98.75	17.388
American Indian/ Alaska Native, non-Hispanic	143	49.83	8.636	149	63.61	11.452	97	72.26	13.008	129	90.53	16.614	81	95.23	15.104
Two or more races, non-Hispanic	722	55.33	12.108	783	70.65	15.460	211	78.26	17.463	682	96.58	18.169	181	104.10	18.475
School type (fall kindergarten)															
Public school	13,581	53.49	11.147	13,529	68.17	14.112	4,187	75.84	15.953	12,022	93.92	17.694	3,840	100.25	17.209
Private school	2,088	56.76	11.425	2,060	71.77	15.486	480	85.35	18.055	1,640	99.00	17.505	415	108.36	16.954

See notes at end of table.

Table 5-21. Reading assessment scale score statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016—Continued

	Spring second grade (X6RSCALK5)			Spring third grade (X7RSCALK5)			Spring fourth grade (X8RSCALK5)			Spring fifth grade (X9RSCALK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**13,837**	**111.93**	**16.922**	**12,866**	**120.66**	**15.331**	**12,074**	**129.31**	**14.513**	**11,427**	**136.26**	**15.337**
Sex												
Male	7,038	110.14	17.602	6,569	119.11	15.985	6,150	128.33	15.513	5,839	135.40	16.274
Female	6,780	113.82	15.954	6,280	122.30	14.429	5,909	130.36	13.295	5,575	137.17	14.221
Race/ethnicity												
White, non-Hispanic	6,534	115.32	16.013	6,092	124.12	14.580	5,705	132.45	13.434	5,421	139.59	13.948
Black, non-Hispanic	1,577	106.91	16.328	1,413	114.59	14.170	1,280	123.49	14.009	1,115	130.01	15.753
Hispanic	3,739	106.41	17.191	3,527	115.70	15.393	3,369	124.87	14.883	3,252	131.66	15.809
Asian, non-Hispanic	1,194	116.34	15.100	1,111	124.50	13.604	1,045	133.71	13.510	1,002	140.74	13.626
Hawaiian, Other Pacific Islander, non-Hispanic	78	112.75	16.257	72	119.16	13.838	67	128.97	13.639	62	136.41	13.423
American Indian/ Alaska Native, non-Hispanic	114	107.26	16.974	106	117.06	15.229	105	126.42	14.720	101	132.89	16.526
Two or more races, non-Hispanic	590	114.90	17.210	535	123.55	15.184	494	131.56	14.872	465	138.29	15.615
School type (fall kindergarten)												
Public school	11,063	111.16	16.888	10,315	119.95	15.423	9,661	128.65	14.523	9,146	135.60	15.453
Private school	1,468	117.91	15.351	1,353	126.10	13.189	1,276	134.67	12.725	1,199	141.43	12.925

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimate (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the number of cases with a valid score. 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 2010, spring 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

5.3.5 Reliabilities

Table 5-22 presents the reliability statistics for the reading assessment scores. The reliabilities shown in table 5-22 are typical and adequate for tests with these numbers of items. The number of items listed in the table is the number of unique items fielded in reading at that round.

Table 5-22. Reading assessment reliabilities for IRT-based scores, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

Score	Number of items	*n*	Weighted reliability
Fall kindergarten	56	15,669	.92
Spring kindergarten	56	17,186	.94
Fall first grade	62	5,194	.95
Spring first grade	62	15,115	.95
Fall second grade	39	4,725	.91
Spring second grade	39	13,837	.90
Spring third grade	24	12,866	.86
Spring fourth grade	27	12,074	.87
Spring fifth grade	21	11,427	.86

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 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 2010, spring 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

5.4 Mathematics Assessment

5.4.1 Samples and Associated Statistics for the Third-Grade Round[6]

The third-grade mathematics assessment consisted of 17 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 25, 20, and 17 items, respectively. Table 5-23 shows the total number of children administered the mathematics assessment and the assessment's associated statistics for the spring third-grade round. No significant ceiling or floor effects were observed in the spring. Classical item analysis results for the mathematics administration showed one item with an

[6] For samples and associated statistics for the kindergarten, first-grade, and second-grade rounds, see the *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a) and the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b).

r-biserial lower than ideal since it was a very difficult item for the sample. There was no Spanish administration of the mathematics assessment in third grade.

Table 5-23. Third-grade mathematics assessment sample, ECLS-K:2011 spring third-grade data collection: Spring 2014

Characteristics	Spring third grade	
	Number	Percent
Total sample size	**12,897**	**100**
Number of children with responses to fewer than 10 items	31	#
Number of children with scoreable data	12,866	100
Number and percent of children with responses only for the router	0	#
Number and percent of children routed to low form	4,073	32
Number and percent of children routed to middle form	5,818	45
Number and percent of children routed to high form	2,974	23
Number and percent of children with a perfect score: router + high form	0	#
Number and percent of children with chance score or below: router + low form	49	#

Rounds to zero.
NOTE: The unweighted *n* is the 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), spring 2014.

5.4.2 Samples and Associated Statistics for the Fourth-Grade Round

The fourth-grade mathematics assessment consisted of 17 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 20, 18, and 16 items, respectively. Table 5-24 shows the total number of children administered the mathematics assessment and the assessment's associated statistics for the spring fourth-grade round. No significant ceiling or floor effects were observed in the spring. Classical item analysis results for the English mathematics administration showed one item with an *r*-biserial slightly lower than ideal because it was a very easy item for the sample. There was no Spanish administration of the mathematics assessment in fourth grade.

Table 5-24. Fourth-grade mathematics assessment sample, ECLS-K:2011 spring fourth-grade data collection: Spring 2015

Characteristics	Spring fourth grade	
	Number	Percent
Total sample size	**12,103**	**100**
Number of children with responses to fewer than 10 items	23	#
Number of children with scoreable data	12,080	100
Number and percent of children with responses only for the router	0	#
Number and percent of children routed to low form	4,231	35
Number and percent of children routed to middle form	6,501	54
Number and percent of children routed to high form	1,347	11
Number and percent of children with a perfect score: router + high form	0	#
Number and percent of children with chance score or below: router + low form	29	#

Rounds to zero.
NOTE: The unweighted *n* is the 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), spring 2015.

5.4.3 Samples and Associated Statistics for the Fifth-Grade Round

The fifth-grade mathematics assessment consisted of 18 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 18, 17, and 16 items, respectively. Table 5-25 shows the total number of children administered the mathematics assessment and the assessment's associated statistics for the spring fifth-grade round. No significant ceiling or floor effects were observed in the spring. Classical item analysis results for the mathematics administration showed no items with *r*-biserials lower than ideal. There was no Spanish administration of the mathematics assessment in fifth grade.

Table 5-25. Fifth-grade mathematics assessment sample, ECLS-K:2011 spring fifth-grade data collection: Spring 2016

Characteristics	Spring fifth grade	
	Number	Percent
Total sample size	**11,445**	**100**
Number of children with responses to fewer than 10 items	19	#
Number of children with scoreable data	11,426	100
Number and percent of children with responses only for the router	0	#
Number and percent of children routed to low form	3,245	28
Number and percent of children routed to middle form	5,299	46
Number and percent of children routed to high form	2,882	25
Number and percent of children with a perfect score: router + high form	3	#
Number and percent of children with chance score or below: router + low form	21	#

Rounds to zero.
NOTE: The unweighted *n* is the 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); spring 2016.

5.4.4 Score Statistics

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 159 unique items administered in the kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade rounds. Plots of the IRT theta and scale score distributions are provided in appendix D. Tables 5-27 and 5-28 present summary statistics by selected child characteristics for the IRT-based mathematics theta and scale scores in the kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade rounds.

Table 5-26. Mathematics assessment statistics by IRT-based score, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

Variable	Description	*n*	Range of possible values	Weighted mean	Weighted standard deviation
X1MTHETK5	X1 MATH THETA-K5	15,595	-4.0–+4.0	-1.15	0.702
X2MTHETK5	X2 MATH THETA-K5	17,143	-4.0–+4.0	-0.40	0.626
X3MTHETK5	X3 MATH THETA-K5	5,222	-4.0–+4.0	-0.03	0.594
X4MTHETK5	X4 MATH THETA-K5	15,103	-4.0–+4.0	0.51	0.554
X5MTHETK5	X5 MATH THETA-K5	4,729	-4.0–+4.0	0.68	0.523
X6MTHETK5	X6 MATH THETA-K5	13,830	-4.0–+4.0	1.04	0.528
X7MTHETK5	X7 MATH THETA-K5	12,866	-4.0–+4.0	1.42	0.462
X8MTHETK5	X8 MATH THETA-K5	12,080	-4.0–+4.0	1.64	0.464
X9MTHETK5	X9 MATH THETA-K5	11,426	-4.0–+4.0	1.83	0.464
X1MSETHK5	X1 MATH STD ERR OF THETA-K5	15,595	0.0–3.0	0.28	0.055
X2MSETHK5	X2 MATH STD ERR OF THETA-K5	17,143	0.0–3.0	0.24	0.036
X3MSETHK5	X3 MATH STD ERR OF THETA-K5	5,222	0.0–3.0	0.23	0.037
X4MSETHK5	X4 MATH STD ERR OF THETA-K5	15,103	0.0–3.0	0.21	0.030
X5MSETHK5	X5 MATH STD ERR OF THETA-K5	4,729	0.0–3.0	0.20	0.032
X6MSETHK5	X6 MATH STD ERR OF THETA-K5	13,830	0.0–3.0	0.19	0.023
X7MSETHK5	X7 MATH STD ERR OF THETA-K5	12,866	0.0–3.0	0.20	0.022
X8MSETHK5	X8 MATH STD ERR OF THETA-K5	12,080	0.0–3.0	0.21	0.034
X9MSETHK5	X9 MATH STD ERR OF THETA-K5	11,426	0.0–3.0	0.19	0.038
X1MSCALK5	X1 MATH IRT SCALE SCORE-K5	15,595	0.0–159.0	35.21	11.479
X2MSCALK5	X2 MATH IRT SCALE SCORE-K5	17,143	0.0–159.0	49.42	13.342
X3MSCALK5	X3 MATH IRT SCALE SCORE-K5	5,222	0.0–159.0	58.01	14.110
X4MSCALK5	X4 MATH IRT SCALE SCORE-K5	15,103	0.0–159.0	72.25	15.500
X5MSCALK5	X5 MATH IRT SCALE SCORE-K5	4,729	0.0–159.0	77.41	15.950
X6MSCALK5	X6 MATH IRT SCALE SCORE-K5	13,830	0.0–159.0	89.72	17.920
X7MSCALK5	X7 MATH IRT SCALE SCORE-K5	12,866	0.0–159.0	103.70	17.802
X8MSCALK5	X8 MATH IRT SCALE SCORE-K5	12,080	0.0–159.0	112.30	17.631
X9MSCALK5	X9 MATH IRT SCALE SCORE-K5	11,426	0.0–159.0	119.45	17.339

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the 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, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 5-27. Math assessment theta statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

	Fall kindergarten (X1MTHETK5)			Spring kindergarten (X2MTHETK5)			Fall first grade (X3MTHETK5)			Spring first grade (X4MTHETK5)			Fall second grade (X5MTHETK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**15,595**	**-1.15**	**0.702**	**17,143**	**-0.40**	**0.626**	**5,222**	**-0.03**	**0.594**	**15,103**	**0.51**	**0.554**	**4,729**	**0.68**	**0.523**
Sex															
Male	7,934	-1.15	0.733	8,720	-0.40	0.658	2,725	-0.04	0.621	7,692	0.52	0.597	2,455	0.71	0.557
Female	7,625	-1.15	0.666	8,392	-0.39	0.591	2,494	-0.02	0.565	7,389	0.49	0.504	2,273	0.65	0.482
Race/ethnicity															
White, non-Hispanic	7,585	-0.97	0.643	8,067	-0.25	0.585	1,934	0.09	0.587	7,093	0.64	0.527	1,754	0.81	0.501
Black, non-Hispanic	2,100	-1.36	0.643	2,223	-0.64	0.610	537	-0.26	0.565	1,801	0.26	0.528	461	0.43	0.470
Hispanic	3,795	-1.46	0.707	4,359	-0.62	0.623	2,004	-0.18	0.549	3,982	0.32	0.530	1,860	0.49	0.497
Asian, non-Hispanic	1,134	-0.82	0.668	1,414	-0.18	0.590	402	0.15	0.589	1,302	0.69	0.506	364	0.86	0.488
Hawaiian, Other Pacific Islander, non-Hispanic	86	-1.28	0.705	113	-0.45	0.626	29	-0.14	0.579	99	0.47	0.494	23	0.73	0.475
American Indian/ Alaska Native, non-Hispanic	143	-1.41	0.765	149	-0.52	0.577	97	-0.14	0.493	129	0.44	0.494	81	0.53	0.434
Two or more races, non-Hispanic	716	-1.05	0.701	781	-0.31	0.589	211	0.00	0.642	682	0.55	0.575	181	0.74	0.521
School type (fall kindergarten)															
Public school	13,516	-1.18	0.706	13,491	-0.43	0.629	4,207	-0.06	0.589	12,010	0.49	0.555	3,844	0.65	0.522
Private school	2,079	-0.90	0.611	2,060	-0.19	0.564	479	0.20	0.540	1,640	0.64	0.508	415	0.84	0.457

See notes at end of table.

Table 5-27. Math assessment theta statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016—Continued

	Spring second grade (X6MTHETK5)			Spring third grade (X7MTHETK5)			Spring fourth grade (X8MTHETK5)			Spring fifth grade (X9MTHETK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**13,830**	**1.04**	**0.528**	**12,866**	**1.42**	**0.462**	**12,080**	**1.64**	**0.464**	**11,426**	**1.83**	**0.464**
Sex												
Male	7,035	1.07	0.558	6,573	1.46	0.480	6,153	1.68	0.482	5,838	1.85	0.477
Female	6,776	1.01	0.492	6,276	1.38	0.437	5,912	1.60	0.440	5,575	1.81	0.449
Race/ethnicity												
White, non-Hispanic	6,531	1.17	0.489	6,091	1.55	0.415	5,706	1.77	0.419	5,421	1.96	0.423
Black, non-Hispanic	1,575	0.74	0.508	1,410	1.13	0.453	1,281	1.35	0.472	1,114	1.55	0.469
Hispanic	3,737	0.87	0.505	3,529	1.27	0.454	3,372	1.50	0.447	3,252	1.69	0.443
Asian, non-Hispanic	1,194	1.27	0.475	1,112	1.61	0.410	1,046	1.83	0.421	1,002	2.04	0.403
Hawaiian, Other Pacific Islander, non-Hispanic	78	1.07	0.494	73	1.38	0.429	67	1.61	0.408	62	1.85	0.360
American Indian/ Alaska Native, non-Hispanic	114	0.94	0.492	106	1.37	0.448	105	1.63	0.409	101	1.80	0.380
Two or more races, non-Hispanic	590	1.12	0.557	535	1.48	0.456	494	1.68	0.478	465	1.84	0.482
School type (fall kindergarten)												
Public school	11,056	1.02	0.529	10,314	1.41	0.466	9,665	1.63	0.468	9,144	1.81	0.470
Private school	1,468	1.18	0.477	1,355	1.52	0.418	1,277	1.76	0.408	1,199	1.95	0.392

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimate (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the number of cases with a valid score. 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 2010, spring 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 5-28. Math assessment scale score statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

	Fall kindergarten (X1MSCALK5)			Spring kindergarten (X2MSCALK5)			Fall first grade (X3MSCALK5)			Spring first grade (X4MSCALK5)			Fall second grade (X5MSCALK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**15,595**	**35.21**	**11.479**	**17,143**	**49.42**	**13.342**	**5,222**	**58.01**	**14.110**	**15,103**	**72.25**	**15.500**	**4,729**	**77.41**	**15.950**
Sex															
Male	7,934	35.38	12.070	8,720	49.46	14.071	2,725	57.88	14.824	7,692	72.93	16.817	2,455	78.51	17.143
Female	7,625	35.05	10.811	8,392	49.41	12.519	2,494	58.15	13.306	7,389	71.54	13.931	2,273	76.25	14.494
Race/ethnicity															
White, non-Hispanic	7,585	37.94	11.401	8,067	52.62	13.092	1,934	61.05	14.356	7,093	76.25	15.271	1,754	81.68	15.798
Black, non-Hispanic	2,100	31.62	9.497	2,223	44.23	11.905	537	52.27	12.369	1,801	65.18	13.189	461	69.57	12.896
Hispanic	3,795	30.49	10.092	4,359	44.67	12.211	2,004	54.21	12.433	3,982	66.79	13.956	1,860	71.59	14.154
Asian, non-Hispanic	1,134	40.94	12.663	1,414	54.24	13.475	402	62.59	14.168	1,302	77.51	15.365	364	83.16	16.313
Hawaiian, Other Pacific Islander, non-Hispanic	86	33.14	10.291	113	48.41	13.285	29	55.30	13.268	99	70.83	13.715	23	79.00	15.143
American Indian/ Alaska Native, non-Hispanic	143	31.41	10.967	149	46.53	12.365	97	54.92	11.908	129	70.16	14.369	81	72.41	13.381
Two or more races, non-Hispanic	716	36.80	11.824	781	51.16	13.328	211	58.89	14.895	682	73.83	16.363	181	79.49	16.157
School type (fall kindergarten)															
Public school	13,516	34.71	11.403	13,491	48.85	13.289	4,207	57.25	13.902	12,010	71.80	15.441	3,844	76.62	15.790
Private school	2,079	39.19	11.305	2,060	53.96	12.891	479	63.63	13.454	1,640	76.14	15.048	415	82.46	14.990

See notes at end of table.

Table 5-28. Math assessment scale score statistics, by child characteristics, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016—Continued

	Spring second grade (X6MSCALK5)			Spring third grade (X7MSCALK5)			Spring fourth grade (X8MSCALK5)			Spring fifth grade (X9MSCALK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**13,830**	**89.72**	**17.920**	**12,866**	**103.70**	**17.802**	**12,080**	**112.30**	**17.631**	**11,426**	**119.45**	**17.339**
Sex												
Male	7,035	91.05	18.883	6,573	105.36	18.433	6,153	113.90	18.125	5,838	120.20	17.768
Female	6,776	88.31	16.725	6,276	101.95	16.932	5,912	110.61	16.930	5,575	118.65	16.836
Race/ethnicity												
White, non-Hispanic	6,531	94.37	17.005	6,091	108.67	16.060	5,706	117.12	15.801	5,421	124.16	15.468
Black, non-Hispanic	1,575	79.40	15.799	1,410	92.46	17.009	1,281	101.17	17.742	1,114	108.60	17.763
Hispanic	3,737	83.64	16.579	3,529	97.81	17.375	3,372	106.78	17.142	3,252	114.00	16.960
Asian, non-Hispanic	1,194	98.00	17.119	1,112	110.88	16.127	1,046	119.44	15.643	1,002	127.09	14.729
Hawaiian, Other Pacific Islander, non-Hispanic	78	90.77	17.024	73	101.85	16.882	67	111.15	15.810	62	120.51	13.687
American Indian/ Alaska Native, non-Hispanic	114	86.04	17.174	106	101.38	17.579	105	111.79	15.718	101	118.45	14.523
Two or more races, non-Hispanic	590	92.78	18.675	535	105.88	17.814	494	113.66	18.254	465	119.73	18.180
School type (fall kindergarten)												
Public school	11,056	89.06	17.864	10,314	103.12	17.925	9,665	111.68	17.745	9,144	118.81	17.566
Private school	1,468	94.42	16.957	1,355	107.62	16.308	1,277	116.74	15.786	1,199	123.83	14.707

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimate (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the number of cases with a valid score. 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 2010, spring 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

5.4.5 Reliabilities

Table 5-29 presents reliability statistics for the scores of the fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade mathematics assessments (calculated in the same way as the reading reliability statistics, described in section 5.3.5). The reliabilities shown in table 5-29 are typical and adequate for tests with this number of items. The number of items listed in the table is the number of unique items fielded in mathematics at that round.

Table 5-29. Mathematics assessment reliabilities, ECLS-K:2011 fall and spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

Score	Number of items	*n*	Weighted reliability
Fall kindergarten	74	15,695	.92
Spring kindergarten	74	17,143	.93
Fall first grade	79	5,222	.93
Spring first grade	79	15,103	.93
Fall second grade	74	4,729	.93
Spring second grade	74	13,830	.94
Spring third grade	62	12,866	.92
Spring fourth grade	55	12,080	.91
Spring fifth grade	57	11,426	.92

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 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 2010, spring 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

5.5 Science Assessment

5.5.1 Samples and Associated Statistics in the Third-Grade Round[7]

The third-grade science assessment consisted of 14 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty) of 14, 16, and 17 items, respectively. Table 5-30 shows the total number of children administered the science assessment and the assessment's associated

[7] For samples and associated statistics for the kindergarten, first-grade, and second-grade rounds, see the *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a) and the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b).

statistics for the spring third-grade round. No significant ceiling or floor effects were observed in the fall or spring. Classical item analysis results for the science administrations showed one very difficult item with an *r*-biserial lower than ideal.

Table 5-30. Third-grade science assessment sample, ECLS-K:2011 spring third-grade data collection: Spring 2014

Characteristics	Spring third grade	
	Number	Percent
Total sample size	**12,897**	**100**
Number of children with responses to fewer than 10 items	41	#
Number of children with scoreable data	12,856	100
Number and percent of children with responses only for the router	0	0
Number and percent of children routed to low form	2,917	23
Number and percent of children routed to middle form	6,531	51
Number and percent of children routed to high form	3,407	27
Number and percent of children with a perfect score: router + high form	0	#
Number and percent of children with chance score or below: router + low form	214	2

Rounds to zero.
NOTE: The unweighted *n* is the 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), spring 2014.

5.5.2 Samples and Associated Statistics in the Fourth-Grade Round

The fourth-grade science assessment consisted of 15 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty), each of which had 13 items. Table 5-31 shows the total number of children administered the science assessment and the assessment's associated statistics for the spring fourth-grade round. No significant ceiling or floor effects were observed in the spring. Classical item analysis results for the science administrations showed one very difficult item with an *r*-biserial lower than ideal.

Table 5-31. Fourth-grade science assessment sample, ECLS-K:2011 spring fourth-grade data collection: Spring 2015

Characteristics	Spring fourth grade	
	Number	Percent
Total sample size	**12,103**	**100**
Number of children with responses to fewer than 10 items	34	#
Number of children with scoreable data	12,069	100
Number and percent of children with responses only for the router	1	#
Number and percent of children routed to low form	1,797	15
Number and percent of children routed to middle form	5,721	47
Number and percent of children routed to high form	4,550	38
Number and percent of children with a perfect score: router + high form	0	#
Number and percent of children with chance score or below: router + low form	30	#

Rounds to zero.
NOTE: The unweighted *n* is the 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), spring 2015.

5.5.3 Samples and Associated Statistics in the Fifth-Grade Round

The fifth-grade science assessment consisted of 15 routing items, followed by one of three second-stage forms (low-, middle-, and high-difficulty), each of which had 13 items. Table 5-32 shows the total number of children administered the science assessment and the assessment's associated statistics for the spring fifth-grade round. No significant ceiling or floor effects were observed in the fall or spring. Classical item analysis results for the science administrations showed no items with an *r*-biserial lower than ideal.

Table 5-32. Fifth-grade science assessment sample, ECLS-K:2011 spring fifth-grade data collection: Spring 2016

Characteristics	Spring fifth grade	
	Number	Percent
Total sample size	**11,445**	**100**
Number of children with responses to fewer than 10 items	26	#
Number of children with scoreable data	11,419	100
Number and percent of children with responses only for the router	1	**#**
Number and percent of children routed to low form	2,441	21
Number and percent of children routed to middle form	4,802	42
Number and percent of children routed to high form	4,175	37
Number and percent of children with a perfect score: router + high form	4	#
Number and percent of children with chance score or below: router + low form	34	#

Rounds to zero.
NOTE: The unweighted *n* is the 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), spring 2016.

5.5.4 Score Statistics

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 100 unique items administered in the spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade rounds. Plots of the IRT theta and scale score distributions are provided in appendix D. Tables 5-34 and 5-35 present summary statistics by selected child characteristics for the IRT-based science theta and scale scores in the kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade rounds.

Table 5-33. Science assessment statistics by IRT-based score, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016

Variable	Description	*n*	Range of possible values	Weighted mean	Weighted standard deviation
X2STHETK5	X2 SCIENCE THETA-K5	16,936	-4.0–+4.0	-0.60	0.737
X3STHETK5	X3 SCIENCE THETA-K5	5,180	-4.0–+4.0	-0.32	0.809
X4STHETK5	X4 SCIENCE THETA-K5	15,072	-4.0–+4.0	0.13	0.786
X5STHETK5	X5 SCIENCE THETA-K5	4,724	-4.0–+4.0	0.40	0.750
X6STHETK5	X6 SCIENCE THETA-K5	13,819	-4.0–+4.0	0.75	0.730
X7STHETK5	X7 SCIENCE THETA-K5	12,856	-4.0–+4.0	1.18	0.650
X8STHETK5	X8 SCIENCE THETA-K5	12,069	-4.0–+4.0	1.53	0.620
X9STHETK5	X9 SCIENCE THETA-K5	11,419	-4.0–+4.0	1.87	0.659
X2SSETHK5	X2 SCIENCE STD ERR OF THETA-K5	16,936	0.0–3.0	0.71	0.081
X3SSETHK5	X3 SCIENCE STD ERR OF THETA-K5	5,180	0.0–3.0	0.51	0.073
X4SSETHK5	X4 SCIENCE STD ERR OF THETA-K5	15,072	0.0–3.0	0.48	0.059
X5SSETHK5	X5 SCIENCE STD ERR OF THETA-K5	4,724	0.0–3.0	0.45	0.065
X6SSETHK5	X6 SCIENCE STD ERR OF THETA-K5	13,819	0.0–3.0	0.43	0.051
X7SSETHK5	X7 SCIENCE STD ERR OF THETA-K5	12,856	0.0–3.0	0.39	0.084
X8SSETHK5	X8 SCIENCE STD ERR OF THETA-K5	12,069	0.0–3.0	0.40	0.067
X9SSETHK5	X9 SCIENCE STD ERR OF THETA-K5	11,419	0.0–3.0	0.37	0.083
X2SSCALK5	X2 SCIENCE IRT SCALE SCORE-K5	16,936	0.0–100.0	33.57	7.353
X3SSCALK5	X3 SCIENCE IRT SCALE SCORE-K5	5,180	0.0–100.0	36.95	9.044
X4SSCALK5	X4 SCIENCE IRT SCALE SCORE-K5	15,072	0.0–100.0	42.71	10.213
X5SSCALK5	X5 SCIENCE IRT SCALE SCORE-K5	4,724	0.0–100.0	46.63	10.722
X6SSCALK5	X6 SCIENCE IRT SCALE SCORE-K5	13,819	0.0–100.0	52.25	11.606
X7SSCALK5	X7 SCIENCE IRT SCALE SCORE-K5	12,856	0.0–100.0	59.83	11.914
X8SSCALK5	X8 SCIENCE IRT SCALE SCORE-K5	12,069	0.0–100.0	66.73	11.902
X9SSCALK5	X9 SCIENCE IRT SCALE SCORE-K5	11,419	0.0–100.0	73.38	12.743

NOTE: Spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the 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), spring 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 5-34. Science assessment theta statistics, by child characteristics, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016

Characteristic	Spring kindergarten (X2STHETK5)			Fall first grade (X3STHETK5)			Spring first grade (X4STHETK5)			Fall second grade (X5STHETK5)		
	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**16,936**	**-0.60**	**0.737**	**5,180**	**-0.32**	**0.809**	**15,072**	**0.13**	**0.786**	**4,724**	**0.40**	**0.750**
Sex												
Male	8,608	-0.59	0.745	2,704	-0.32	0.813	7,676	0.15	0.800	2,451	0.43	0.767
Female	8,299	-0.61	0.729	2,473	-0.32	0.805	7,374	0.10	0.770	2,272	0.37	0.730
Race/ethnicity												
White, non-Hispanic	8,045	-0.31	0.600	1,929	-0.02	0.649	7,088	0.39	0.653	1,751	0.65	0.592
Black, non-Hispanic	2,217	-0.93	0.709	537	-0.71	0.756	1,800	-0.23	0.735	461	0.10	0.680
Hispanic	4,198	-1.01	0.748	1,969	-0.76	0.856	3,962	-0.25	0.848	1,860	0.00	0.852
Asian, non-Hispanic	1,402	-0.77	0.738	400	-0.39	0.822	1,299	0.09	0.791	363	0.40	0.731
Hawaiian, Other Pacific Islander, non-Hispanic	113	-0.96	0.822	29	-0.90	0.697	99	-0.07	0.823	23	0.43	0.602
American Indian/ Alaska Native, non-Hispanic	146	-0.63	0.747	97	-0.21	0.773	129	0.17	0.688	81	0.34	0.630
Two or more races, non-Hispanic	778	-0.40	0.627	211	-0.11	0.754	680	0.32	0.733	180	0.57	0.785
School type (fall kindergarten)												
Public school	13,329	-0.63	0.745	4,179	-0.36	0.805	11,988	0.10	0.792	3,841	0.37	0.753
Private school	2,060	-0.35	0.620	479	0.02	0.718	1,639	0.37	0.660	415	0.67	0.626

See notes at end of table.

Table 5-34. Science assessment theta statistics, by child characteristics, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016 —Continued

Characteristic	Spring second grade (X6STHETK5)			Spring third grade (X7STHETK5)			Spring fourth grade (X8STHETK5)			Spring fifth grade (X9STHETK5)		
	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**13,819**	**0.75**	**0.730**	**12,856**	**1.18**	**0.650**	**12,069**	**1.53**	**0.620**	**11,419**	**1.87**	**0.659**
Sex												
Male	7,026	0.78	0.744	6,563	1.20	0.664	6,146	1.55	0.625	5,835	1.89	0.669
Female	6,774	0.72	0.712	6,276	1.15	0.634	5,908	1.50	0.613	5,571	1.85	0.648
Race/ethnicity												
White, non-Hispanic	6,524	0.97	0.615	6,085	1.37	0.559	5,704	1.71	0.528	5,420	2.05	0.557
Black, non-Hispanic	1,572	0.39	0.694	1,409	0.83	0.635	1,276	1.19	0.639	1,110	1.48	0.701
Hispanic	3,737	0.44	0.784	3,528	0.93	0.686	3,370	1.31	0.644	3,252	1.66	0.677
Asian, non-Hispanic	1,193	0.85	0.748	1,112	1.32	0.583	1,045	1.63	0.586	1,001	2.02	0.608
Hawaiian, Other Pacific Islander, non-Hispanic	78	0.56	0.848	73	0.96	0.669	66	1.32	0.658	62	1.81	0.615
American Indian/ Alaska Native, non-Hispanic	114	0.71	0.639	106	1.08	0.567	105	1.53	0.552	100	1.81	0.621
Two or more races, non-Hispanic	590	0.91	0.715	533	1.32	0.646	494	1.63	0.631	465	1.93	0.745
School type (fall kindergarten)												
Public school	11,047	0.72	0.734	10,306	1.15	0.655	9,657	1.51	0.621	9,138	1.84	0.665
Private school	1,467	0.96	0.623	1,352	1.39	0.566	1,276	1.72	0.542	1,198	2.06	0.558

NOTE: Spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the number of cases with a valid score. 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 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 5-35. Science assessment scale score statistics, by child characteristics, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016

	Spring kindergarten (X2SSCALK5)			Fall first grade (X3SSCALK5)			Spring first grade (X4SSCALK5)			Fall second grade (X5SSCALK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**16,936**	**33.57**	**7.353**	**5,180**	**36.95**	**9.044**	**15,072**	**42.71**	**10.213**	**4,724**	**46.63**	**10.722**
Sex												
Male	8,608	33.68	7.457	2,704	36.98	9.145	7,676	43.05	10.383	2,451	47.17	10.943
Female	8,299	33.46	7.241	2,473	36.92	8.936	7,374	42.36	10.019	2,272	46.06	10.455
Race/ethnicity												
White, non-Hispanic	8,045	36.39	6.637	1,929	40.20	8.272	7,088	46.10	9.356	1,751	50.21	9.685
Black, non-Hispanic	2,217	30.21	6.358	537	32.50	7.316	1,800	37.86	8.660	461	42.01	9.218
Hispanic	4,198	29.57	6.685	1,969	32.25	8.291	3,962	38.01	9.823	1,860	41.24	10.419
Asian, non-Hispanic	1,402	31.79	7.018	400	36.24	9.316	1,299	42.21	10.374	363	46.48	10.698
Hawaiian, Other Pacific Islander, non-Hispanic	113	30.26	7.265	29	30.40	6.484	99	40.20	9.669	23	46.65	9.539
American Indian/ Alaska Native, non-Hispanic	146	33.26	7.095	97	38.13	8.725	129	43.01	8.826	81	45.32	9.087
Two or more races, non-Hispanic	778	35.40	6.765	211	39.43	9.033	680	45.36	10.051	180	49.49	10.871
School type (fall kindergarten)												
Public school	13,329	33.26	7.369	4,179	36.50	8.873	11,988	42.41	10.205	3,841	46.16	10.598
Private school	2,060	36.03	6.735	479	41.01	8.694	1,639	45.81	9.496	415	50.58	10.064

See notes at end of table.

Table 5-35. Science assessment scale score statistics, by child characteristics, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016—Continued

	Spring second grade (X6SSCALK5)			Spring third grade (X7SSCALK5)			Spring fourth grade (X8SSCALK5)			Spring fifth grade (X9SSCALK5)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**13,819**	**52.25**	**11.606**	**12,856**	**59.83**	**11.914**	**12,069**	**66.73**	**11.902**	**11,419**	**73.38**	**12.743**
Sex												
Male	7,026	52.77	11.810	6,563	60.40	12.164	6,146	67.20	12.001	5,835	73.78	12.856
Female	6,774	51.70	11.359	6,276	59.22	11.613	5,908	66.24	11.776	5,571	72.96	12.608
Race/ethnicity												
White, non-Hispanic	6,524	55.75	10.545	6,085	63.37	10.689	5,704	70.26	10.418	5,420	76.97	10.798
Black, non-Hispanic	1,572	46.24	10.234	1,409	53.21	10.900	1,276	60.00	11.660	1,110	65.70	13.364
Hispanic	3,737	47.43	11.364	3,528	55.20	11.946	3,370	62.34	12.060	3,252	69.24	13.152
Asian, non-Hispanic	1,193	54.08	12.016	1,112	62.38	11.051	1,045	68.79	11.278	1,001	76.42	11.710
Hawaiian, Other Pacific Islander, non-Hispanic	78	49.58	12.000	73	55.68	11.641	66	62.63	12.329	62	72.26	12.063
American Indian/ Alaska Native, non-Hispanic	114	51.20	10.643	106	57.69	10.481	105	66.66	10.976	100	72.29	11.998
Two or more races, non-Hispanic	590	55.00	11.541	533	62.66	12.078	494	68.72	12.086	465	74.81	13.813
School type (fall kindergarten)												
Public school	11,047	51.81	11.578	10,306	59.30	11.918	9,657	66.25	11.915	9,138	72.92	12.850
Private school	1,467	55.69	10.820	1,352	63.76	10.958	1,276	70.41	10.635	1,198	77.10	10.898

NOTE: Spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. See the ECLS-K:2011 data file User's Manuals for an explanation of the weighting variables. The unweighted *n* is the number of cases with a valid score. 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 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

5.5.5 Reliabilities

Table 5-36 presents reliability statistics for the spring kindergarten, fall and spring first-grade, fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-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 acceptable based on the intended application of the ECLS-K:2011 test scores (Weiner, Schinka, and Velicer 2004). The number of items listed in the table is the number of unique items fielded in science at that round.

Table 5-36. Science assessment reliabilities, ECLS-K:2011 spring kindergarten, fall and spring first-grade, fall and spring second-grade, third-grade, fourth-grade, and fifth-grade data collections: Spring 2011, school year 2011–12, school year 2012–13, spring 2014, spring 2015, and spring 2016

Score	Number of items	*n*	Weighted reliability
Spring kindergarten	20	16,936	.73
Fall first grade	43	5,180	.83
Spring first grade	43	15,072	.84
Fall second grade	50	4,724	.86
Spring second grade	50	13,819	.85
Spring third grade	47	12,856	.83
Spring fourth grade	40	12,069	.82
Spring fifth grade	40	11,419	.86

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 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 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

5.6 Evaluating the Kindergarten Through Fifth-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 fifth 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)?

5.6.1 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 5.5). 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.

5.6.2 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, second grade, third grade, fourth grade, and fifth 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, and on each spring round in the third-grade, fourth-grade, and fifth-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.

5.6.3 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 nonunidimensionality. 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, second-grade, third-grade, fourth-grade, and fifth-grade assessments, there was a total of 26 unique polytomous and 64 unique dichotomous items resulting in a total of 167 score points contributing to the scoring model for the reading assessment. Of the 90 calibrated items in the reading scale (26 polytomous and 64 dichotomous), 70 have *a* parameter estimates greater than 1.0. Items with *a* parameter estimates slightly below 1.0 are predominantly related to polytomous item sets associated with passages, while those with the lowest *a* parameter estimates are associated with some of the easiest items in the assessment.

Results for mathematics were similar. Across the kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade assessments, there was a total of 159 unique items contributing to the scoring model for the mathematics assessment, with 136 of the 159 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, second-grade, third-grade, fourth-grade, and fifth-grade assessments, there was a total of 100 unique items contributing to the scoring model for the science assessment. In the science assessment, 61 of the 100 items had *a* parameter estimates above 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 addition to examining the *a* parameters in support of one underlying factor, differential item functioning (DIF) contrasting data from the fall and spring in kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade 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).

5.6.3.1 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.

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

Table 5-37 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 5-37, 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.

Table 5-37. Correlations of IRT theta score across rounds, by subject, ECLS-K:2011: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Subject	Fall kindergarten	Spring kindergarten	Fall first grade	Spring first grade	Fall second grade	Spring second grade	Spring third grade	Spring fourth grade	Spring fifth grade
Reading									
Fall kindergarten	1.00	†	†	†	†	†	†	†	†
Spring kindergarten	0.67	1.00	†	†	†	†	†	†	†
Fall first grade	0.54	0.65	1.00	†	†	†	†	†	†
Spring first grade	0.58	0.71	0.60	1.00	†	†	†	†	†
Fall second grade	0.45	0.60	0.51	0.65	1.00	†	†	†	†
Spring second grade	0.53	0.67	0.59	0.77	0.75	1.00	†	†	†
Spring third grade	0.44	0.52	0.42	0.58	0.62	0.69	1.00	†	†
Spring fourth grade	0.41	0.52	0.43	0.58	0.65	0.66	0.65	1.00	†
Spring fifth grade	0.42	0.53	0.44	0.57	0.63	0.64	0.63	0.73	1.00
Mathematics									
Fall kindergarten	1.00	†	†	†	†	†	†	†	†
Spring kindergarten	0.66	1.00	†	†	†	†	†	†	†
Fall first grade	0.56	0.68	1.00	†	†	†	†	†	†
Spring first grade	0.56	0.68	0.71	1.00	†	†	†	†	†
Fall second grade	0.52	0.65	0.74	0.79	1.00	†	†	†	†
Spring second grade	0.54	0.66	0.76	0.81	0.84	1.00	†	†	†
Spring third grade	0.48	0.59	0.62	0.68	0.74	0.76	1.00	†	†
Spring fourth grade	0.46	0.58	0.66	0.70	0.74	0.77	0.74	1.00	†
Spring fifth grade	0.44	0.55	0.62	0.67	0.72	0.74	0.74	0.78	1.00

See notes at end of table.

Table 5-37. Correlations of IRT theta score across rounds, by subject, ECLS-K:2011: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Subject	Fall kindergarten	Spring kindergarten	Fall first grade	Spring first grade	Fall second grade	Spring second grade	Spring third grade	Spring fourth grade	Spring fifth grade
Science									
Spring kindergarten	†	1.00	†	†	†	†	†	†	†
Fall first grade	†	0.69	1.00	†	†	†	†	†	†
Spring first grade	†	0.62	0.74	1.00	†	†	†	†	†
Fall second grade	†	0.59	0.73	0.76	1.00	†	†	†	†
Spring second grade	†	0.57	0.72	0.77	0.81	1.00	†	†	†
Spring third grade	†	0.50	0.61	0.69	0.74	0.75	1.00	†	†
Spring fourth grade	†	0.46	0.57	0.64	0.70	0.73	0.74	1.00	†
Spring fifth grade	†	0.43	0.55	0.60	0.68	0.68	0.69	0.74	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.

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, spring 2013, spring 2014, spring 2015, and spring 2016.

Correlations of scores *across* subjects *within* rounds are presented in table 5-38. 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 Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) and Early Childhood Longitudinal Study, Birth Cohort (ECLS-B).

Table 5-38. Correlations of IRT theta score across subjects, by round, ECLS-K:2011: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

Round	Reading and mathematics	Reading and science	Mathematics and science
Fall kindergarten	0.70	†	†
Spring kindergarten	0.73	0.57	0.58
Fall first grade	0.53	0.74	0.59
Spring first grade	0.77	0.64	0.70
Fall second grade	0.71	0.58	0.71
Spring second grade	0.75	0.65	0.74
Spring third grade	0.65	0.64	0.77
Spring fourth grade	0.72	0.67	0.76
Spring fifth grade	0.75	0.66	0.76

† 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, spring 2013, spring 2014, spring 2015, and spring 2016.

5.7 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.

5.7.1 Choosing the Appropriate Score for Analysis

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

- The IRT-based theta scores are overall measures of ability. They are appropriate for both cross-sectional and longitudinal analyses. They are useful in examining differences in overall achievement among subgroups of children in a given data collection round or across rounds, as well as in analysis looking at correlations between achievement and child, family, and school characteristics. 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.[8] However, for a broader audience of readers unfamiliar with IRT modeling techniques, the metric of the theta scores (from -8 to 8) 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.

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

An important issue to be considered when analyzing achievement scores and gains is assessment timing: children's age at assessment, the date of assessment, and the time interval between assessments. Most sampled children were born throughout the second half of 2004 and first half of 2005, but their birth dates were not related to testing dates. As a result, children were tested at different developmental and chronological ages. Assessment dates ranged from August to December for the fall data 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

[8] 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.

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 passages), 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 child to learn in the classroom. Two measures of executive function were included in fall and spring of kindergarten, first grade, and second grade, and in the spring of third, fourth, and fifth grades: 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* (WJ III) *Tests of Cognitive Abilities* (Woodcock, McGrew, and Mather 2001), assessing working memory. The same version of the DCCS was administered in fall and spring of the kindergarten year and fall and spring of first grade. In second grade, the DCCS was changed to computerized administration to remain age appropriate through fifth grade. The same computerized version was used in fall and spring of second grade and in the spring of third grade, fourth grade, and fifth grade. The Numbers Reversed task remained the same across all rounds of collection, kindergarten through fifth grade. In fourth grade, a third measure of executive function was added. The National Institutes of Health (NIH) Toolbox Flanker Inhibitory Control and Attention Task (Flanker) task (Zelazo et al. 2013), which measures inhibitory control in the context of selective visual attention, was administered for the first time in fourth grade, and then the same version of the task was administered again in the fifth grade.

6.1 Dimensional Change Card Sort

The DCCS (Zelazo 2006; Zelazo et al. 2013) was used to assess 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. Beginning with 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. The shift to a computerized version of the task was made so that the DCCS would remain age appropriate through the end of data collection for the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011). For more information on the physical, table-top card sort task administered in kindergarten and first grade and differences between the physical version and computerized version, see chapter 3 of the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES

2017-285) (Tourangeau et al. 2017). This section describes the computerized version of the DCCS that was administered in the second-grade, third-grade, fourth-grade, and fifth-grade rounds.

The computerized task was developed as part of the NIH Toolbox and is appropriate for ages 3–85 (Weintraub et al. 2013; Zelazo et al. 2013). The task had been under development during the planning phases for the earliest rounds of the ECLS-K:2011 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) is a task that is used across the 3 through 85 age range, but it has two different start points based on the age of the child in order to limit administration time. The NIH Toolbox DCCS consists of 40 trials, including 5 pre-switch trials (where 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). Testing conducted in the development of the NIH Toolbox DCCS indicated 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 above begin with the mixed-block trials and are given credit in the scoring for completing the pre-switch and post-switch trials accurately.

For the ECLS-K:2011 administrations of the computerized DCCS, all ECLS-K:2011 children were administered the version of the NIH Toolbox DCCS for ages 8 years and older, regardless of their age at the time of assessment. In second grade, approximately 90 percent of the ECLS-K:2011 children in the fall subsample for second grade and approximately 40 percent of children in the spring of second grade who had a score on the DCCS were not yet 8 years old when the DCCS was administered. In third grade, nearly all children who participated in the DCCS (99.95 percent) were at least 8 years old when the DCCS was administered. In fourth and fifth grades, all children who participated in the DCCS were at least 8 years old when the DCCS was administered. The decision to administer the same version of the DCCS from second grade forward, regardless of whether the child was age 8, was made so that all study children would receive the same version of the DCCS task in second grade and in later rounds of data collection. Use of the same measure allows for a longitudinal analysis of performance on the DCCS from second grade into later rounds of data collection.

As noted earlier, the construct assessed in the physical version of the DCCS that was administered in kindergarten and first grade 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 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.

In each of the 30 mixed-block trials administered via computer to children in the ECLS-K:2011 beginning in the second-grade rounds, 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 the screen. Next, the stimulus picture was displayed in the center of the 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 the 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. 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 sort by either 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). 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 between rules 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. 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 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 administered only 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. 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.

Item-level data for the 30 test 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 (C*DCCS1-C*DCCS30); (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) (C*GAME1-C*GAME30); and (3) reaction times reported in milliseconds (C*TARGRT1-C*TARGRT30). Variable names for the

item-level data from the fall and spring second-grade assessments begin with "C5" and "C6," respectively. Item-level variable names from the assessments in spring third grade, spring fourth grade, and spring fifth grade begin with "C7", "C8," and "C9," respectively.

The overall computed score reported for the second-, third-, fourth-, and fifth-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). Scores 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. Accuracy is considered first. If the child's accuracy rate is less than or equal to 80 percent, the child's overall computed score is based entirely on accuracy. If the child's accuracy rate is more than 80 percent, the child's overall computed score is based on a combination of accuracy and reaction time.

The accuracy score factored into the computation of the overall score can range from 0 to 5. There is a total of 40 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 five pre-switch and the five post-switch trials of the DCCS that were not administered. Therefore, the accuracy component of the overall computed DCCS score is calculated as follows:

DCCS accuracy score = 0.125 * number of correct responses.[1]

If the child's accuracy rate is higher than 80 percent, a reaction time score is added to the child's accuracy score.[2] Like the accuracy score, the reaction time score ranges from 0 to 5 points.

The reaction time component of the overall 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

[1] The number of correct responses = 10 + the number of correct trials out of the 30 mixed-block trials. Once the child has passed the practice trials and advanced into the scored portion of the assessment, 10 accuracy points are automatically awarded due to the chosen start point for the task. For this reason, 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.

[2] 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-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).

frequently used sorting rule, color), following the same scoring algorithm outlined in the scoring manual for the NIH Toolbox (Slotkin, Nowinski et al. 2012). First, for those children with greater than 80 percent accuracy on the 40 trials, the median reaction time is calculated based on reaction times for correct nondominant trials with reaction times greater than or equal to 100 milliseconds (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 were 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.[3]

To summarize, the overall computed score on the computerized DCCS 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 overall 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).

The fall and spring second-grade, spring third-grade, spring fourth-grade, and spring fifth-grade computed scores (X5DCCSSCR, X6DCCSSCR, X7DCCSSCR, X8DCCSSCR, and X9DCCSSCR) 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 score. The overall computed score 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

[3] 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 reaction time was not factored into the child's score.

directly compared to performance in the spring of second grade, the spring of third grade, the spring of fourth grade, and the spring of fifth 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. For example, approximately 4 percent of children in the third-grade data collection who have a computed overall score failed to achieve greater than 80 percent accuracy. In fourth grade, this percentage was approximately 2 percent. In fifth grade, approximately 1 percent of children who have a computed overall score did not achieve greater than 80 percent accuracy. The score for these children is calculated based solely on accuracy. The remaining children (96 percent in third grade, 98 percent in fourth grade, and 99 percent in fifth grade) who have a computed overall score 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.

6.1.1 Mean Scores for the Dimensional Change Card Sort

This section presents the overall means for the DCCS scores from the computerized administration in spring second, spring third, spring fourth, and spring fifth grades. Means for the overall DCCS score by data collection round and by selected child characteristics are also presented.

Five scores based on the computerized administration of the DCCS are presented on the data file for fall and spring of second grade, spring of third grade, spring of fourth grade, and spring of fifth

grade:[4] overall score (X*DCCSSCR; range: 0-10); accuracy score (X*CSACC; range: 0-5) that is scaled as described above to compute the overall DCCS score; reaction time score (X*CSNDRT; range: 0-5) that is scaled to compute the overall DCCS score; count of correct, dominant trials (X*CSDAC; range: 0-23); and count of correct nondominant trials (X*CSNDAC; range: 0-7). Researchers should note that the count of correct dominant trials and the count of correct nondominant trials represent accuracy by trial type for the 30 administered trials and are different from the total accuracy score (X*CSACC, DCCS Accuracy Component [0-5] Score) that is derived to compute the overall DCCS computed score. Researchers should also note that the reaction time score was only computed for cases for which the accuracy score was greater than 80 percent. If the accuracy score was not greater than 80 percent, then the reaction time score was set to -9 (not ascertained).

Table 6-1 presents the variable names, descriptions, value ranges, weighted means, and standard deviations for the second-, third-, fourth-, and fifth-grade DCCS scores. For information on the kindergarten and first-grade scores, see the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES 2017-285) (Tourangeau et al. 2017).

Table 6-2 presents the mean overall DCCS scores for spring second, spring third, spring fourth, and spring fifth grades by data collection round and by selected child characteristics.

[4] An asterisk "*" is a placeholder for the round number in variable names. Fall kindergarten and spring kindergarten are rounds 1 and 2, respectively. Fall first grade and spring first grade are rounds 3 and 4, respectively. Fall second grade and spring second grade are rounds 5 and 6, respectively. Spring third grade is round 7, spring fourth grade is round 8, and spring fifth grade is round 9.

Table 6-1. Dimensional Change Card Sort variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School year 2012–13, spring 2014, spring 2015, and spring 2016

Variable name	Description	*n*	Range of possible values[1]	Weighted mean	Standard deviation
X5DCCSSCR	X5 Computed (Overall) Score	4,708	0–10	6.37	1.402
X6DCCSSCR	X6 Computed (Overall) Score	13,774	0–10	6.69	1.345
X7DCCSSCR	X7 Computed (Overall) Score	12,744	0–10	7.19	1.098
X8DCCSSCR	X8 Computed (Overall) Score	12,021	0–10	7.63	0.965
X9DCCSSCR	X9 Computed (Overall) Score	11,386	0–10	7.97	0.943
X5CSACC	X5 DCCS Accuracy Component (0–5) Score	4,708	0–5	4.53	0.589
X6CSACC	X6 DCCS Accuracy Component (0–5) Score	13,774	0–5	4.59	0.504
X7CSACC	X7 DCCS Accuracy Component (0–5) Score	12,744	0–5	4.72	0.356
X8CSACC	X8 DCCS Accuracy Component (0–5) Score	12,021	0–5	4.80	0.274
X9CSACC	X9 DCCS Accuracy Component (0–5) Score	11,386	0–5	4.82	0.246
X5CSNDRT	X5 DCCS Nondom RT Component (0–5) Score	4,067	0–5	2.09	0.758
X6CSNDRT	X6 DCCS Nondom RT Component (0–5) Score	12,405	0–5	2.33	0.765
X7CSNDRT	X7 DCCS Nondom RT Component (0–5) Score	12,222	0–5	2.58	0.777
X8CSNDRT	X8 DCCS Nondom RT Component (0–5) Score	11,790	0–5	2.88	0.768
X9CSNDRT	X9 DCCS Nondom RT Component (0–5) Score	11,247	0–5	3.19	0.790
X5CSDAC	X5 DCCS Dominant Trial Accuracy Count	4,708	0–23	20.19	4.468
X6CSDAC	X6 DCCS Dominant Trial Accuracy Count	13,774	0–23	20.62	3.758
X7CSDAC	X7 DCCS Dominant Trial Accuracy Count	12,744	0–23	21.53	2.535
X8CSDAC	X8 DCCS Dominant Trial Accuracy Count	12,021	0–23	22.05	1.852
X9CSDAC	X9 DCCS Dominant Trial Accuracy Count	11,386	0–23	22.18	1.638
X5CSNDAC	X5 DCCS Nondominant Trial Accuracy Count	4,708	0–7	6.08	1.128
X6CSNDAC	X6 DCCS Nondominant Trial Accuracy Count	13,774	0–7	6.11	1.100
X7CSNDAC	X7 DCCS Nondominant Trial Accuracy Count	12,744	0–7	6.21	1.011
X8CSNDAC	X8 DCCS Nondominant Trial Accuracy Count	12,021	0–7	6.33	0.926
X9CSNDAC	X9 DCCS Nondominant Trial Accuracy Count	11,386	0–7	6.40	0.865

[1] 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 obtain an accuracy score of 0. Therefore, the lowest accuracy component (0-5) score in the data file is 1.25**,** and the lowest computed (overall) score in the data file is also 1.25.

NOTE: Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.

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, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 6-2. Mean Dimensional Change Card Sort total score for spring second grade, spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2013, spring 2014, spring 2015, and spring 2016

	Spring second grade (X6DCCSSCR)			Spring third grade (X7DCCSSCR)			Spring fourth grade (X8DCCSSCR)			Spring fifth grade (X9DCCSSCR)		
Characteristic	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Total sample	**13,774**	**6.69**	**1.345**	**12,744**	**7.19**	**1.098**	**12,021**	**7.63**	**0.965**	**11,386**	**7.97**	**0.943**
Sex												
Male	7,003	6.60	1.424	6,509	7.17	1.158	6,119	7.62	1.007	5,816	7.96	0.993
Female	6,752	6.79	1.249	6,218	7.22	1.030	5,887	7.64	0.918	5,557	7.98	0.887
Race/ethnicity												
White, non-Hispanic	6,505	6.85	1.228	6,041	7.31	0.996	5,684	7.71	0.907	5,409	8.05	0.886
Black, non-Hispanic	1,567	6.16	1.630	1,388	6.81	1.353	1,277	7.35	1.166	1,109	7.74	1.096
Hispanic	3,722	6.59	1.358	3,496	7.10	1.103	3,360	7.56	0.937	3,237	7.88	0.834
Asian, non-Hispanic	1,191	6.78	1.325	1,109	7.35	1.054	1,032	7.79	0.906	997	8.14	0.976
Hawaiian, Other Pacific Islander, non-Hispanic	78	6.49	1.528	72	7.10	1.054	67	7.36	1.108	62	8.12	0.835
American Indian/Alaska Native, non-Hispanic	114	7.04	1.112	104	7.41	1.078	102	7.95	0.781	100	8.17	0.877
Two or more races, non-Hispanic	588	6.88	1.201	524	7.27	1.109	489	7.68	0.994	463	8.00	1.173
School type												
Public school	12,461	6.67	1.355	11,549	7.18	1.108	10,905	7.62	0.968	10,326	7.96	0.948
Private school	1,273	6.88	1.244	1,151	7.36	0.982	1,057	7.68	0.944	1,002	8.06	0.890

NOTE: Spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. Detail may not sum to totals 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 2013, spring 2014, spring 2015, and spring 2016.

6.1.2 Use of Dimensional Change Card Sort Scores in Longitudinal Analyses

Two different versions of the DCCS were administered, as described above in section 6.1. A physical version of the task was administered in kindergarten and first grade (rounds 1–4), and a computerized version of the task was administered in second through fifth grades (rounds 5–9). Although both tasks measured cognitive flexibility, the way the construct was assessed and the scoring differed across the versions. This makes longitudinal comparisons between the two versions of the task more difficult. Performance in rounds 1–4 can be directly compared because the same task was used in each of these rounds. Performance in rounds 5–9 can be directly compared because the same task was used in each of

these rounds. However, performance in rounds 1–4 cannot be directly compared to performance in rounds 5–9. The scores for rounds 1–4 were calculated differently and are on a different metric than the scores for rounds 5–9. As described in more detail below, the total score in rounds 1–4 (X*DCCSTOT) ranges from 0 to 18 and is based solely on the accuracy of sorting the cards, whereas the total score in rounds 5–9 (X*DCCSSCR) ranges from 0 to 10 and is based on both accuracy and reaction time.

The total score in kindergarten through first grade (rounds 1–4) is a raw score that is the sum of the number of correct trials in the set of six pre-switch trials (Color Game), the set of six post-switch trials (Shape Game), and the set of six border trials (Border Game). The score ranges from 0 to 18. All children were administered the six pre-switch trials and the six post-switch trials. The border trials were administered if the child got at least five of the post-switch trials (Shape Game) correct. In the spring of first grade, the mean total score (X4DCCSTOT) was around 16 out of a possible score of 18, which indicates that some children reached ceiling performance on the physical card sort at first grade. In fall kindergarten, 7.7 percent of children had the maximum score (18) on the physical card sort, and in spring of kindergarten, 15.8 percent of children had the maximum score (i.e., at ceiling). In fall first grade, 23.5 percent of children had a score of 18, and in spring first grade 29.3 percent of children had reached ceiling on the physical version of the DCCS. Zelazo (2006) found that by age 5 most children were able to switch sorting rules when asked to do so and, therefore, it was not surprising that many children performed at ceiling in first grade. However, having a large percentage of children at ceiling underestimates performance for these children. DCCS performance in rounds 1 through 4 can be directly compared. Comparison among these rounds indicates raw improvement or decline between rounds.

The total score in the second- through fifth-grade data collections (rounds 5–9) is a computed score that ranges from 0 to 10. A score between 0 and 5 indicates that the child had lower accuracy (80 percent correct or less) and did not trade speed for accuracy (i.e., did not slow down as necessary in order to get the trial correct). This means that accuracy rather than reaction time is the better indicator of performance. A score between 5 and 10 indicates accuracy was high enough to factor reaction time into the score. When a child is developmentally able to trade speed for accuracy and appropriately slow down reaction time in order to be accurate, then reaction time becomes the critical aspect of the scoring and is factored into the scoring equation. The total score on the computerized DCCS (X*DCCSSCR) in second through fifth grades (rounds 5–9) can be directly compared. Comparison between these rounds indicates raw improvement or decline between rounds. A change in the child's score from one round to another represents real change in the level of performance for that child.

Data users need to carefully consider whether and how to compare performance across the physical and computerized versions of the task. Although the tasks are conceptually the same and the

computerized version extends the difficulty level to make it age appropriate for older ages, the numbers of trials administered and the scoring is different, so scores cannot be directly compared. One possible strategy is to standardize each of the total scores so that each is on the same scale, with the same mean and standard deviation (e.g., create *z* scores).

The ECLS-K:2011 does not provide age-adjusted or grade-adjusted scale scores. Computation of such scores is left to the discretion of the data user, based on the research question.

Additionally, conceptually the pre-switch (Color Game) and post-switch (Shape Game) trials in the physical version of the DCCS are the same as the pre-switch and post-switch trials in the computerized version of the task. Although the physical version of the DCCS used red or blue rabbits or boats and had physical cards that were sorted into piles, and the computerized version of the DCCS used yellow or blue balls or trucks that were sorted electronically with key presses, the pre-switch and post-switch trials were the same. Whereas the physical version of the DCCS used border trials to increase the difficulty of the task for young children, the computerized version of the DCCS used mixed-block trials to increase the difficulty. Reaction time was not recorded for the physical version of the DCCS, but reaction time was an integral part of the scoring for the computerized version of the task when accuracy was sufficiently high (over 80 percent).

6.1.3 Dimensional Change Card Sort Data Flags

Nine flags indicate the presence or absence of DCCS data. X1DCCSFLG and X2DCCSFLG indicate the presence of data for the fall and spring of kindergarten, respectively. X3DCCSFLG and X4DCCSFLG indicate the presence of first-grade data for the fall and spring, respectively; X5DCCSFLG and X6DCCSFLG indicate that data are present for the overall computed DCCS score (X5DCCSSCR/X6DCCSSCR) for the fall and spring of second grade, respectively; X7DCCSFLG indicates that data are present for the overall computed DCCS score (X7DCCSSCR) for the spring of third grade; X8DCCSFLG indicates that data are present for the overall DCCS score (X8DCCSSCR) for the spring of fourth grade; and X9DCCSFLG indicates that data are present for the overall DCCS score (X9DCCSSCR) for the spring of fifth grade.

The use of computers for the administration of the DCCS in second, third, fourth, and fifth grades allowed the completion flags (X5DCCSFLG, X6DCCSFLG, X7DCCSFLG, X8DCCSFLG, X9DCCSFLG) to be developed with additional detail that was not available for kindergarten and first grade. The values indicate whether the task was administered, whether the overall computed DCCS score is

present, and, if a score is not present, the reason why it is 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 (meaning 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 (for example, a fire drill or the child needing to return to class) that interrupted an assessment session. Note that the Shape Game preceded the Color Game during the practice trials. There are differences between the second-grade, third-grade, fourth-grade, and fifth-grade DCCS flags, as explained below.

The DCCS flags for the fall and spring of second grade, the spring of fourth grade, and the spring of fifth grade have six possible values. A description of the values of these completion flags is presented in exhibit 6-1.

Exhibit 6-1. Data flag description for the computerized Dimensional Change Card Sort for fall and spring second grade, spring fourth grade, and spring fifth grade: School year 2012–13, spring 2015, and spring 2016

X*DCCSFLG[1]	Value
Not Administered	0
DCCS computed (overall) score present	1
Failed Shape Game practice	2
Failed Color Game practice	3
Breakoff before passing practice trials	4
Breakoff after passing practice trials	5

[1]An asterisk "*" is a placeholder for the round number in variable names. The data flag variables for fall (round 5) and spring (round 6) second grade, spring fourth grade (round 8), and spring fifth grade (round 9) are X5DCCSFLG, X6DCCSFLG, X8DCCSFLG, and X9DCCSFLG, respectively.
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, spring 2013, spring 2015, and spring 2016.

The DCCS flag variable for the spring of third grade, X7DCCSFLG, ranges from 0 to 7. A description of the values of the completion flag is presented in exhibit 6-2. Two additional codes not used in second, fourth, and fifth grade were added to the third-grade flag to identify a small number of cases that were affected by a programming error that occurred in the third-grade administration of the DCCS. This error resulted in giving children credit for a correct response when the child did not provide a response to a trial. This scoring error occurred in both the practice and test trials. Scoring errors that occurred during the test trials were corrected in the data. These errors did not affect the child's experience during the test, but only affected how the trial was recorded.

Exhibit 6-2. Data flag description for the computerized Dimensional Change Card Sort for spring third grade: Spring 2014

X7DCCSFLG	Value
Not Administered	0
DCCS computed (overall) score present	1
Failed Shape Game practice	2
Failed Color Game practice	3
Breakoff before passing practice trials	4
Breakoff after passing practice trials	5
Programming error but still passed practice, DCCS data present	6
Programming error, insufficient practice, DCCS data set to -4	7

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014.

Errors that occurred during the third-grade practice trials, however, did affect the child's experience during the test and, in some cases, resulted in insufficient opportunity for the child to demonstrate an understanding of the rules of the game. When a child did not respond to a trial in the practice, the program treated the nonresponse as a correct response and provided incorrect audio feedback to the child. The audio feedback that the child heard was "That's right," even though the child did not provide a response. If the child did not respond to a trial, the trial was supposed to be scored as incorrect, and the audio feedback was supposed to indicate that the child responded with an incorrect answer and reteach the rule. The erroneous feedback during the practice could have confused the child about the rules of the game. It is important for the child to demonstrate a clear understanding of the rules of the game in the practice trials before progressing to the test trials to ensure that performance is not a reflection of failing to understand the instructions. Under some circumstances, having nonresponse scored as correct affected what practice trials were administered.

Cases affected by the third-grade programming error were examined to determine whether they met the criteria for moving into the test trials based on the items for which they did provide a response (that is, whether they demonstrated sufficient understanding of the task despite receiving erroneous feedback). These cases, children who had at least one instance of nonresponse in the practice, are flagged as a 6 or 7 in the DCCS flag variable depending on whether they met the criteria. Cases that have X7DCCSFLG = 6 passed the practice trials with the responses they provided during the administration of the DCCS. For example, a child may have had three correct responses and one nonresponse within the block of four practice trials and, thus, the criterion of responding correctly to at least three of four correct in order to proceed was still reached. As another example, the child could have had two nonresponse trials and two incorrect trials and failed the first practice set. In this case, the child would have been administered another practice block of four trials and could have passed on that set of practice trials. Cases that have the value of 6 on the DCCS flag are cases that successfully met the criteria for passing both the shape and color

practice and advanced to the test trial, despite receiving at least one instance of erroneous feedback. There are 189 cases that have X7DCCSFLG = 6, and data for these cases are provided on the data file. Additional information on this error is provided in the appendix of the data file user's manual.

Cases that have X7DCCSFLG = 7 did not demonstrate sufficient understanding of the task with the responses they provided and were not given sufficient practice per the administration protocols to have their scores included in the data file. These cases were not given the opportunity to meet the criterion for passing the practice because nonresponse was incorrectly recorded as a correct response. For example, children who had two correct trials, one incorrect trial and one nonresponse trial (incorrectly scored as "correct") were incorrectly given credit for passing the practice, even though they only had two correct trials and did not meet the criterion of at least three of four correct to pass. In this example, if the program had performed correctly, the child would have been given additional training and additional opportunities to pass the practice. Because of the programming error, this did not happen and the child progressed to the test trials without truly meeting the criterion for successfully passing the practice. Because it was not possible to determine whether the children could have passed the practice if given the correct opportunities, the data were suppressed. There are 92 cases that have X7DCCSFLG = 7. These cases have DCCS data set to -4 (suppressed due to insufficient practice).

6.2 Numbers Reversed

The Numbers Reversed 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 three consecutive two-number sequences incorrect, then the Numbers Reversed task ends. If the child does not get three consecutive two-number sequences incorrect, the child is then given up to 5 three-number sequences. The sequence becomes increasingly longer, up to a maximum of eight numbers, until the child gets three consecutive number sequences of the same length incorrect (or completes all number sequences).

Item-level data for the Numbers Reversed subtask for the fall and spring of kindergarten, first grade, and second grade and spring of third grade, fourth grade, and fifth grade are provided in the ECLS-K:2011 K-5 data file. The maximum number of items any child could have been administered in all data collection rounds 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 from 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, his or her case would have a missing code for the Numbers Reversed scores. Variable names for the item-level data from the fall kindergarten assessments begin with "C1," and variable names for the item-level data from the spring kindergarten assessments begin with "C2." Similarly, variable names for item-level data from the fall and spring first-grade assessments begin with "C3" and "C4," while those for fall and spring second grade and spring third grade begin with "C5," "C6," and "C7," respectively. Variable names for the item-level data from the spring fourth-grade assessment begin with "C8," and variable names for the item-level data from the spring fifth-grade assessment begin with "C9." Variable descriptions for these items indicate the length of the digit sequence (e.g., C1 Numbers Reversed Two-digit sequence #1). In addition to the item-level data, five scores developed using guidelines from the publisher's scoring materials are included in the data file for Numbers Reversed: the *W*-ability[5] score, the age standard score, the grade standard score, the age percentile score, and the grade percentile score.

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.[6]

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 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 of 500, and most older children would be expected to have scores above the mean of 500. Also, as a child develops 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

[5] The *W*-ability score is a *W* score that represents the individual's level of ability on the task presented.
[6] More information on these publisher scores can be found in the Technical Manual for the *Woodcock-Johnson III* (McGrew and Woodcock 2001).

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 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.

In kindergarten and first grade, the Numbers Reversed subtask was administered in both English and Spanish. It was administered in Spanish to children routed through the assessment battery in Spanish because they did not pass an English-language screener.[7] 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; nevertheless, 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 a raw score of 0). While this difference in the *W* scores between English and Spanish administrations is largest at the lower end of the *W* distribution, the difference occurs along the entirety of the *W* distribution. For example, a raw score of 11 corresponds to a *W* score of 496 in the English administration norming data and a *W* score of 494 in the Spanish administration norming data. The data file includes one *W* score variable per round of data collection that contains data for all children administered the Numbers Reversed task, regardless of the language of administration. Researchers who want to account for language of administration in their analyses can use the data flag provided on the data file for each round (X*FLSCRN) to identify which children were administered Numbers Reversed in English and which children were administered Numbers Reversed in Spanish. All children were

[7] 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 of the *ECLS-K:2011 Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a).

administered the assessments in English starting with the second-grade data collection. Therefore, the second-, third-, fourth-, and fifth-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 through fifth grades are not provided on the data file.

Although the *W* score is reflective of the average performance of 10-year-olds, and the ECLS-K:2011 children are younger in the earlier rounds of the study, it is included in the data file to enable the measurement of changes in children's working memory longitudinally across all rounds of the study. Also, it facilitates 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 primarily 5 or 6 years old during the kindergarten data collections, 6 or 7 years old during the first-grade data collections, 7 or 8 years old during the second-grade data collections, 8 or 9 years old during the third-grade data collection, 9 and 10 years old during the fourth-grade data collection, and 10 and 11 years old during the fifth-grade data collection[8] 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 from kindergarten through fourth grade and above average performance in fifth grade. However, because the mean of the *W* scale was set by the publisher based on the average performance for a child 10 years, 0 months, this pattern is as expected. As expected, the discrepancy declined as the participating children grew older and closer to age 10. Because the average age at assessment was approximately age 11 years in the spring of fifth grade, it is not surprising that the average W score is above 500, the mean set for the average performance of a child 10 years, 0 months.

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 measurable by the WJ III Numbers Reversed task, the study is able to compare children's baseline Numbers Reversed *W* score (fall kindergarten and/or spring kindergarten Numbers Reversed *W* score) with children's 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 getting at least one answer correct.

In the fall of kindergarten, approximately 40 percent of students did not demonstrate sufficient skills as measured by this assessment to score above the lowest scalable score (403 for English assessment and 393 for Spanish assessment). In the spring of kindergarten, approximately 20 percent of students did

[8] For the fourth-grade assessment, approximately 56 percent of the children were 10 years old or older, and approximately 44 percent of the children were 9 years old or younger. For the fifth-grade assessment, nearly all the children were 10 years old or older (99.9 percent).

not score above the lowest scalable score (403 for English, 393 for Spanish). In the fall of first grade, less than 13 percent scored at the lowest scalable score, and only 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. In the spring of third grade, approximately 1 percent scored at the lowest scalable score. In the spring of fourth grade, 0.6 percent scored at the lowest scalable score. In the spring of fifth grade, 0.5 percent scored the lowest scalable score.

A factor that may contribute to the large number of children scoring 403 (and 393 for Spanish) 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 determine the extent to which improper administration of the practice items affected the results. However, readers should keep in mind that this may have affected performance for some (but not all) children. In conducting analyses, 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. For the first-grade and later data collections, assessor training for the Numbers Reversed task was changed to improve the consistency and clarity of administration of the practice items. The instructions trainers provided to the assessors emphasized the need to present practice items consistently and to present multiple practice items when necessary. More information about the Numbers Reversed scoring and data can be found in the *ECLS-K:2011 Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a).

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 0 to 100 relative to the performance of subjects in the WJ III norming sample that is at the same age or grade as the ECLS-K:2011 subjects.

As with the kindergarten and first-grade *W* scores, the kindergarten and first-grade standard scores and percentile scores in the data file contain data from both the English and Spanish administrations of the Numbers Reversed task. Standard scores and percentile scores are a function of the child's age or

grade at assessment. The publisher's scoring protocols result in standard and percentile scores that extend to slightly lower ages for children who were administered the task in Spanish compared to children who were administered the task in English, again due to differences in the precision of measurement within the norming samples. Children 62 months and younger who were administered the Numbers Reversed task in English and who earned a raw score of 0 or 1 have a *W* score but do not have a standard score or percentile score (*W* scores are a function of the number correct and not a function of age). However, all children who were administered this task in Spanish, including those aged 62 months and younger have a *W* score, standard scores, and percentile scores, regardless of their raw score. Again, researchers who want to account for language of administration in their analyses during kindergarten or first grade can use the variables X1FLSCRN, X2FLSCRN, X3FLSCRN, and X4FLSCRN to identify language.

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 for Numbers Reversed

This section presents means for the Numbers Reversed scores. Means for the Numbers Reversed scores by data collection round and by child characteristics are also presented.

Five Numbers Reversed scores were calculated for each round of data collection and are presented on the data file: the *W*-ability score (X*NRWABL), age standard score (X*NRSSCR), grade standard score (X*NRSSGR), age percentile rank (X*NRPERC), and grade percentile rank (X*NRPEGR). The variable names, descriptions, value ranges, weighted means, and standard deviations for the Numbers Reversed scores from the fall of kindergarten to the spring of fifth grade are shown in table 6-3.

The weighted means for the ECLS-K:2011 population are lower than the established means from the WJ III norming sample in some rounds and higher than the established means from the WJ III norming sample in other rounds.[9] For example, the average *W* scores for the ECLS-K:2011 population are less than 500 in kindergarten through fourth grade but higher than 500 in fifth grade. The average age standard scores are less than 100 in all rounds. The average grade standard scores are less than 100 in kindergarten through second grade but higher than 100 in third through fifth grades. The average age and grade percentile scores are less than 50 in some rounds and above 100 in other rounds. 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 (generally 10-year-olds are in fourth or fifth grade)[10] or to differences between the ECLS-K:2011 population and the WJ III norming sample. The differences between weighted means for the average age and grade standard scores and percentile scores for the ECLS-K:2011 population compared to the established means from the WJ III norming sample may also be attributable to differences between the ECLS-K:2011 population and the WJ III norming sample.

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

[10] For the fourth-grade assessment, approximately 56 percent of the children were 10 years old or older, and approximately 44 percent of the children were 9 years old or younger. For the fifth-grade assessment, nearly all children were 10 years old or older (99.9 percent).

Table 6-3. Numbers Reversed variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1NRWABL	X1 Numbers Reversed *W*-Ability Score	15,598	393–603	432.56	30.028
X1NRSSCR	X1 Numbers Reversed Age Standard Score	14,445	45–200	93.10	16.510
X1NRSSGR	X1 Numbers Reversed Grade Standard Score	15,598	33–200	96.40	14.569
X1NRPERC	X1 Numbers Reversed Age Percentile	14,445	0–100	37.89	31.786
X1NRPEGR	X1 Numbers Reversed Grade Percentile	15,598	0–100	41.98	30.886
X2NRWABL	X2 Numbers Reversed *W*-Ability Score	17,147	393–603	449.49	30.412
X2NRSSCR	X2 Numbers Reversed Age Standard Score	17,124	39–200	94.92	17.017
X2NRSSGR	X2 Numbers Reversed Grade Standard Score	17,147	33–200	94.76	16.049
X2NRPERC	X2 Numbers Reversed Age Percentile	17,124	0–100	42.44	30.970
X2NRPEGR	X2 Numbers Reversed Grade Percentile	17,147	0–100	41.89	29.980
X3NRWABL	X3 Numbers Reversed *W*-Ability Score	5,222	393–603	458.42	27.990
X3NRSSCR	X3 Numbers Reversed Age Standard Score	5,221	36–200	94.21	16.969
X3NRSSGR	X3 Numbers Reversed Grade Standard Score	5,222	24–200	95.19	17.815
X3NRPERC	X3 Numbers Reversed Age Percentile	5,221	0–100	41.23	28.832
X3NRPEGR	X3 Numbers Reversed Grade Percentile	5,222	0–100	43.61	29.857
X4NRWABL	X4 Numbers Reversed *W*-Ability Score	15,107	393–603	469.56	25.395
X4NRSSCR	X4 Numbers Reversed Age Standard Score	15,102	24–200	95.90	16.872
X4NRSSGR	X4 Numbers Reversed Grade Standard Score	15,107	19–200	95.42	18.159
X4NRPERC	X4 Numbers Reversed Age Percentile	15,102	0–100	44.35	28.470
X4NRPEGR	X4 Numbers Reversed Grade Percentile	15,107	0–100	44.07	29.276
X5NRWABL	X5 Numbers Reversed *W*-Ability Score	4,727	403–603	473.93	23.736
X5NRSSCR	X5 Numbers Reversed Age Standard Score	4,727	29–200	94.93	16.574
X5NRSSGR	X5 Numbers Reversed Grade Standard Score	4,727	19–200	95.85	17.561
X5NRPERC	X5 Numbers Reversed Age Percentile	4,727	0–100	42.13	27.609
X5NRPEGR	X5 Numbers Reversed Grade Percentile	4,727	0–100	44.17	28.742
X6NRWABL	X6 Numbers Reversed *W*-Ability Score	13,832	403–603	480.70	22.841
X6NRSSCR	X6 Numbers Reversed Age Standard Score	13,828	25–200	95.80	16.749
X6NRSSGR	X6 Numbers Reversed Grade Standard Score	13,832	18–200	95.52	17.715
X6NRPERC	X6 Numbers Reversed Age Percentile	13,828	0–100	43.67	27.765
X6NRPEGR	X6 Numbers Reversed Grade Percentile	13,832	0–100	43.59	28.680
X7NRWABL	X7 Numbers Reversed *W*-Ability Score	12,877	403–603	489.78	21.624
X7NRSSCR	X7 Numbers Reversed Age Standard Score	12,874	20–200	96.34	16.185
X7NRSSGR	X7 Numbers Reversed Grade Standard Score	12,877	18–195	102.74	17.037
X7NRPERC	X7 Numbers Reversed Age Percentile	12,874	0–100	44.10	27.742
X7NRPEGR	X7 Numbers Reversed Grade Percentile	12,877	0–100	55.90	28.907

See notes at end of table.

Table 6-3. Numbers Reversed variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X8NRWABL	X8 Numbers Reversed *W*-Ability Score	12,085	403–603	497.17	21.333
X8NRSSCR	X8 Numbers Reversed Age Standard Score	12,082	15–192	96.65	15.975
X8NRSSGR	X8 Numbers Reversed Grade Standard Score	12,085	19–200	101.86	16.819
X8NRPERC	X8 Numbers Reversed Age Percentile	12,082	0–100	44.28	27.780
X8NRPEGR	X8 Numbers Reversed Grade Percentile	12,085	0–100	54.01	28.724
X9NRWABL	X9 Numbers Reversed *W*-Ability Score	11,430	403–603	503.12	22.005
X9NRSSCR	X9 Numbers Reversed Age Standard Score	11,429	130–182	96.67	16.494
X9NRSSGR	X9 Numbers Reversed Grade Standard Score	11,430	19–200	100.92	17.017
X9NRPERC	X9 Numbers Reversed Age Percentile	11,429	0–100	44.34	28.576
X9NRPEGR	X9 Numbers Reversed Grade Percentile	11,430	0–100	52.28	29.149

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P_30, and spring first-grade estimates (X4) are weighted by W4CS4P_20. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_20. Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.
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, spring 2013, spring 2014, spring 2015, and spring 2016.

Means and standard deviations for the Numbers Reversed scores are provided by data collection round and by child characteristics in tables 6-4 through 6-8. The *W*-ability scores, age standard scores, grade standard scores, age percentile scores, and grade percentile scores are provided by child characteristics for the third-grade, fourth-grade, and fifth-grade data collections. 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* (NCES 2018-182) (Najarian et al. 2018a). The grade standard scores and the grade percentile scores for fall and spring kindergarten are provided by child characteristics in the *ECLS-K:2011 First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b). The fall and spring first-grade and fall and spring second-grade *W*-ability scores, age standard scores, grade standard scores, age percentile scores, and grade percentile scores are provided by child characteristics in the *ECLS-K:2011 First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b).

Table 6-4. Mean Numbers Reversed *W*-ability score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

	Spring third grade (X7NRWABL)			Spring fourth grade (X8NRWABL)			Spring fifth grade (X9NRWABL)		
Characteristic	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Total sample	**12,877**	**489.78**	**21.624**	**12,085**	**497.17**	**21.333**	**11,430**	**503.12**	**22.005**
Sex									
Male	6,579	488.96	21.971	6,159	496.54	21.783	5,841	502.64	22.564
Female	6,281	490.64	21.216	5,911	497.84	20.825	5,576	503.62	21.385
Race/ethnicity									
White, non-Hispanic	6,096	492.17	20.785	5,711	499.32	20.857	5,423	505.33	21.647
Black, non-Hispanic	1,412	483.05	23.198	1,280	491.11	23.142	1,114	497.86	22.546
Hispanic	3,532	487.01	21.067	3,373	494.76	20.292	3,254	500.13	20.768
Asian, non-Hispanic	1,113	495.49	22.775	1,047	503.70	19.774	1,002	510.20	23.002
Hawaiian, Other Pacific Islander, non-Hispanic	73	489.46	18.345	66	498.22	20.847	62	504.28	18.698
American Indian/Alaska Native, non-Hispanic	106	490.53	18.644	105	496.74	18.744	101	501.17	20.543
Two or more races, non-Hispanic	535	491.45	22.850	494	497.28	23.548	465	502.89	25.429
School type									
Public school	11,673	489.37	21.655	10,967	496.73	21.383	10,365	502.78	22.037
Private school	1,159	494.70	19.974	1,059	502.00	20.043	1,007	507.46	20.861

NOTE: Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. Detail may not sum to totals 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 2014, spring 2015, and spring 2016.

Table 6-5. Mean Numbers Reversed age standard score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

	Spring third grade (X7NRSSCR)			Spring fourth grade (X8NRSSCR)			Spring fifth grade (X9NRSSCR)		
Characteristic	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Total sample	**12,874**	**96.34**	**16.185**	**12,082**	**96.65**	**15.975**	**11,429**	**96.67**	**16.494**
Sex									
Male	6,578	95.60	16.470	6,158	96.06	16.351	5,841	96.21	16.960
Female	6,280	97.12	15.841	5,910	97.28	15.543	5,576	97.17	15.970
Race/ethnicity									
White, non-Hispanic	6,094	97.91	15.665	5,709	98.07	15.650	5,423	98.18	16.187
Black, non-Hispanic	1,412	91.47	17.119	1,280	92.23	17.312	1,114	92.84	17.068
Hispanic	3,532	94.54	15.711	3,373	95.10	15.181	3,254	94.67	15.631
Asian, non-Hispanic	1,113	101.35	17.194	1,047	102.09	14.818	1,002	102.35	17.043
Hawaiian, Other Pacific Islander, non-Hispanic	72	95.96	13.719	65	97.24	15.700	61	97.31	14.605
American Indian/Alaska Native, non-Hispanic	106	96.35	13.710	105	96.05	14.262	101	95.05	15.143
Two or more races, non-Hispanic	535	97.62	17.289	494	96.77	17.619	465	96.40	19.139
School type									
Public school	11,671	96.05	16.201	10,965	96.33	16.013	10,364	96.44	16.523
Private school	1,158	99.86	15.064	1,058	100.16	14.966	1,007	99.76	15.540

NOTE: Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. Detail may not sum to totals 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 2014, spring 2015, and spring 2016.

Table 6-6. Mean Numbers Reversed grade standard score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

	Spring third grade (X7NRSSGR)			Spring fourth grade (X8NRSSGR)			Spring fifth grade (X9NRSSGR)		
Characteristic	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Total sample	**12,877**	**102.74**	**17.037**	**12,085**	**101.86**	**16.819**	**11,430**	**100.92**	**17.017**
Sex									
Male	6,579	102.18	17.248	6,159	101.40	17.176	5,841	100.58	17.465
Female	6,281	103.33	16.792	5,911	102.35	16.419	5,576	101.27	16.522
Race/ethnicity									
White, non-Hispanic	6,096	104.61	16.567	5,711	103.50	16.464	5,423	102.59	16.678
Black, non-Hispanic	1,412	97.57	17.700	1,280	97.25	18.157	1,114	96.95	17.558
Hispanic	3,532	100.49	16.477	3,373	99.94	16.016	3,254	98.62	16.110
Asian, non-Hispanic	1,113	107.30	18.171	1,047	106.98	15.639	1,002	106.35	17.855
Hawaiian, Other Pacific Islander, non-Hispanic	73	102.23	14.659	66	102.62	16.775	62	101.64	14.669
American Indian/Alaska Native, non-Hispanic	106	103.32	15.005	105	101.81	14.853	101	99.96	15.757
Two or more races, non-Hispanic	535	104.20	18.072	494	102.06	18.635	465	100.82	19.819
School type									
Public school	11,673	102.44	17.037	10,967	101.51	16.852	10,365	100.66	14.041
Private school	1,159	106.42	16.127	1,059	105.67	15.880	1,007	104.25	16.088

NOTE: Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. Detail may not sum to totals 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 2014, spring 2015, and spring 2016.

Table 6-7. Mean Numbers Reversed age percentile score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

	Spring third grade (X7NRPERC)			Spring fourth grade (X8NRPERC)			Spring fifth grade (X9NRPERC)		
Characteristic	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Total sample	**12,874**	**44.10**	**27.742**	**12,082**	**44.28**	**27.780**	**11,429**	**44.34**	**28.576**
Sex									
Male	6,578	43.06	27.575	6,158	43.44	27.931	5,841	43.69	28.582
Female	6,280	45.19	27.875	5,910	45.16	27.592	5,576	45.03	28.554
Race/ethnicity									
White, non-Hispanic	6,094	46.66	27.669	5,709	46.74	27.762	5,423	47.00	28.578
Black, non-Hispanic	1,412	35.92	27.207	1,280	36.83	27.874	1,114	37.75	28.163
Hispanic	3,532	41.12	26.562	3,373	41.26	26.600	3,254	40.60	27.295
Asian, non-Hispanic	1,113	53.26	28.739	1,047	54.37	27.118	1,002	55.02	28.373
Hawaiian, Other Pacific Islander, non-Hispanic	72	42.74	24.607	65	46.32	24.446	61	45.47	22.880
American Indian/Alaska Native, non-Hispanic	106	43.81	22.585	105	41.74	26.827	101	39.67	27.823
Two or more races, non-Hispanic	535	46.34	29.745	494	45.11	28.339	465	44.37	30.554
School type									
Public school	11,671	43.61	27.626	10,965	43.69	27.724	10,364	43.91	28.533
Private school	1,158	49.89	28.015	1,058	50.68	27.406	1,007	49.91	28.313

NOTE: Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. Detail may not sum to totals 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 2014, spring 2015, and spring 2016.

Table 6-8. Mean Numbers Reversed grade percentile score for spring third grade, spring fourth grade, and spring fifth grade, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

	Spring third grade (X7NRPEGR)			Spring fourth grade (X8NRPEGR)			Spring fifth grade (X9NRPEGR)		
Characteristic	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Total sample	**12,877**	**55.90**	**28.907**	**12,085**	**54.01**	**28.724**	**11,430**	**52.28**	**29.149**
Sex									
Male	6,579	55.21	28.904	6,159	53.38	29.007	5,841	51.84	29.190
Female	6,281	56.63	28.894	5,911	54.67	28.405	5,576	52.74	29.099
Race/ethnicity									
White, non-Hispanic	6,096	58.99	28.302	5,711	56.84	28.324	5,423	55.25	28.761
Black, non-Hispanic	1,412	46.79	29.658	1,280	45.90	29.798	1,114	45.32	29.546
Hispanic	3,532	52.54	28.411	3,373	50.65	27.937	3,254	48.19	28.375
Asian, non-Hispanic	1,113	63.39	28.398	1,047	63.36	26.563	1,002	62.17	27.979
Hawaiian, Other Pacific Islander, non-Hispanic	73	55.56	25.340	66	57.79	25.674	62	55.08	23.208
American Indian/Alaska Native, non-Hispanic	106	58.80	23.423	105	52.73	28.186	101	49.04	28.237
Two or more races, non-Hispanic	535	57.54	30.259	494	54.72	29.270	465	51.93	31.108
School type									
Public school	11,673	55.42	28.868	10,967	53.39	28.749	10,365	51.81	29.145
Private school	1,159	61.80	28.162	1,059	60.69	27.533	1,007	58.25	28.294

NOTE: Spring third-grade estimates (X7) are weighted by W7C7P_20. Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. Detail may not sum to totals 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 2014, spring 2015, and spring 2016.

6.2.2 Use of Numbers Reversed Scores in Longitudinal Analyses

Numbers Reversed was administered in all rounds of data collection (rounds 1–9). The same task was used in each round. There are five scores provided on the data file for each round of data collection, in addition to the item-level data: the *W*-ability score (X*NRWABL), age standard score (X*NRSSCR), grade standard score (X*NRSSGR), age percentile rank (X*NRPERC), and grade percentile rank (X*NRPEGR). These scores were discussed in section 6.2.

Depending on the research question and analysis being conducted, one of the scores may be more preferable than another. 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 *W* score is an equal-interval scale that is useful for measuring growth. For example, a child with a score that increases from 420 to 440 has a 20-point gain, and a child with a score that increases from 500 to 520 has a 20-point gain. These scores show the same amount of growth, although the starting level of ability is different. 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* score for each child in the ECLS-K:2011 was determined using norming data provided by the publisher. Each 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.[11]

Also, as a child develops 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.

6.2.3 Numbers Reversed Data Flags

Nine flags indicate the presence or absence of Numbers Reversed data. X1NRFLG and X2NRFLG indicate the presence of data for the fall and spring of kindergarten, respectively. X3NRFLG and X4NRFLG indicate the presence of first-grade data for the fall and spring, respectively, and X5NRFLG and X6NRFLG indicate the presence of fall and spring second-grade data, respectively. X7NRFLG, X8NRFLG, and X9NRFLG indicate the presence of data for spring third grade, spring fourth grade, and spring fifth grade, respectively.

There is one other flag, X*NRGEST, related to Numbers Reversed that is provided for each round of data collection. The Numbers Reversed grade-normed scores (X*NRSSGR, X*NRPEGR) are normed according to how far into the school year the assessment was conducted. Decimals are used to indicate the number of months into the school year the child had been in the grade at the time of the assessment (e.g., 0.1 = 1 month; 0.2 = 2 months, etc.; 0.9 = 9 months, including time in the summer prior to the start of the next grade level). When school year start and end dates were not available, it was necessary to estimate the decimal representing the proportion of the school year completed when the assessment occurred. X*NRGEST indicates whether the number of months completed in the grade was estimated for

[11] In kindergarten and first grade, the Numbers Reversed subtask was administered in both English and Spanish. In second through fifth grades, Numbers Reversed was administered in English only.

that round of data collection. Time in grade was estimated for approximately 2 percent of children in third and fourth grades and for approximately 3 percent of children in fifth grade.

6.3 The NIH Toolbox Flanker Inhibitory Control and Attention Task (Flanker)

The NIH Toolbox Flanker Inhibitory Control and Attention Task (Flanker) is a computerized task that was developed as part of the NIH Toolbox for the Assessment of Neurological and Behavioral Function (NIH Toolbox) and is appropriate for ages 3–85 (Weintraub et al. 2013; Zelazo et al. 2013). The Flanker was adapted from the Attention Network Test (ANT; e.g., Rueda et al. 2004), which was based on the Eriksen flanker task (Eriksen and Eriksen 1974). It is a measure of executive function; specifically, it is a measure of inhibitory control in the context of selective visual attention (Slotkin, Nowinski, et al. 2012). The Flanker (Zelazo et al. 2013) was added to the ECLS-K:2011 assessment battery in fourth grade, and it was administered again in fifth grade.

The ECLS-K:2011 used the version of the NIH Toolbox Flanker task that is for children 8 years and older.[12] Starting with the fourth-grade administration of the ECLS-K:2011, all children were at least 8 years old. In this task children must inhibit an automatic response tendency that may interfere with achieving a goal and use selective attention to consciously direct sensory or thought processes to a stimulus in the visual field in the service of goal-directed behavior. In the Flanker task, children are asked to focus attention on a central stimulus while ignoring or inhibiting attention to stimuli presented on either side of the central stimulus. The stimulus used for children 8 years and older is a series of five arrows, pointing either left or right. The arrows that "flank" the central arrow, which are referred to as "flankers," either point in the same direction as the central arrow (congruent) or in the direction opposite the central arrow (incongruent). The flanker arrows act as distractors, taking attention away from the central arrow that is supposed to be the focus of the child's attention. Children are presented with 20 arrow trials and are asked to press a button on the computer to indicate the direction the central stimulus (arrow) is pointing. Like the DCCS, the score based on the Flanker is derived from a formula that takes into consideration both accuracy and reaction time (Zelazo et al. 2013; Slotkin, Nowinski, et al. 2012). Performance on the incongruent trials is used to derive a score that is a measure of inhibitory control in the context of selective visual attention.

[12] The NIH Toolbox Flanker task has two different start points based on the age of the child. Children aged 3-7 begin the task with trials that use fish as the stimulus and progress to harder trials that use arrows as stimuli if performance on the fish trials is 90 percent or more correct. By design, children who are 8 years and older begin with the arrow trials and are given credit for successful completion of the fish trials because it was determined that the majority of children 8 years and older could successfully complete the easier fish trials. The task includes two different start points in order to reduce participant burden and create a task with a shorter administration time. Because all children in the ECLS-K:2011 study were at least 8 years of age in the fourth- and fifth-grade data collections, all of them began with the arrow trials and were given credit in the scoring for successfully completing the fish trials.

At the start of the 20 test trials, children were given the instruction to do the task as quickly as possible, and visual and auditory supports were provided to help the children direct attention appropriately. Each of the test trials began with a picture of a star presented on the screen in the location where the central (target) stimulus was about to appear. The star served to direct the child's gaze and orient the child's attention to a standard location, the location where the child needed to be looking. Next, the word "MIDDLE" appeared on the screen in the same location while a prerecorded female voice said "middle," to remind the child to look at the middle arrow and to indicate the direction of that arrow. Next, a series of five arrows appeared on the screen in a line, and the child's task was to press the left arrow key if the arrow in the middle of the five arrows (i.e., the central arrow) was pointing to the left or press the right arrow key if the central arrow was pointing to the right. Children were instructed to "Keep your eyes on the star. Answer as fast as you can without making mistakes. If you make a mistake, just keep going."

The 20 test trials were the same for all children. The direction of the central arrow was counterbalanced across the 20 trials, and there were more congruent trials than incongruent trials. There were 13 congruent trials (central arrow pointed in the same direction as the arrows flanking it) and 7 incongruent trials (central arrow pointed in the opposite direction as the arrows flanking it). For example, the central arrow for trial 1 was left-facing, and the flankers were congruent; the central arrow for trial 2 was right-facing, and the flankers were congruent; and the center arrow for trial 3 was right-facing, and the flankers were incongruent (i.e., left-facing). Like the DCCS, the congruent and incongruent trials in the Flanker were intermixed across the trials, and the number of congruent trials preceding an incongruent trial did not follow a pattern. Congruent trials were more frequent 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). A predisposition to respond based on the orientation of the distractors flanking the central stimulus further increases the difficulty of the incongruent trials; the child must ignore or inhibit attention to the distractors, and this is easier to do when the flankers are congruent. Congruent trials are easier because there is no conflict between the central stimulus and its flankers since all the arrows are pointing in the same direction. Incongruent trials are more difficult because the flankers pointing in the opposition direction from the central stimulus create a distraction with conflicting information. The child needs to respond based solely on the direction of the central stimulus rather than the conflicting and distracting information. To do this, the child must selectively attend to the central arrow, inhibit attention to the conflicting and distracting information provided by the flankers, and inhibit an automatic tendency to respond based on the direction of the flankers.

There is a "cost" in performance that is associated with the conflicting and distracting information presented in the incongruent trials. As discussed in the section on the DCCS, the "cost" to the child's performance on this task that is associated with this conflict can be seen in either more errors or a

slower reaction or response time on incongruent trials. The type of "cost" that is demonstrated (more errors vs. slower reaction time) tends to differ by the age of the participant (Davidson et al. 2006). Younger children tend to demonstrate this cost by having more errors in performance, whereas older children tend to demonstrate this cost by having slower reaction times. Younger children tend to make more errors on incongruent trials because they tend to respond quickly without making an adjustment for the need to ignore the conflict presented by the distractors. Younger children do not slow themselves down in favor of higher accuracy, and, therefore, accuracy is a better metric of performance for young children (Zelazo et al. 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, older children and adults demonstrate their "cost" to ignore the conflict of the incongruent flankers in terms of their reaction time on incongruent trials. Using a scoring method that takes both speed and accuracy into consideration is a strategy for overcoming the challenge of comparing scores of children with developmental differences in the ability to make a speed-accuracy tradeoff. The scoring algorithm used to produce scores from the data collected by the Flanker is analogous to the formula used for the computerized DCCS. The scoring algorithm factors in reaction time on the incongruent trials but only when the child demonstrates sufficiently accurate performance across all the test trials, defined as being accurate on more than 80 percent of the trials (Zelazo et al. 2013). Thus, the Flanker provides a measure of performance through this developmental shift to learning to trade speed for accuracy. More information on scoring is provided below.

The 20 test trials were administered only to children who successfully completed the practice portion of the Flanker. The assessor instructed the child on how to do the task by reading the standardized task instructions that appeared on the screen alongside example stimuli and by familiarizing the child with the response buttons to use on the computer keyboard (left and right arrow key). The child could be presented with up to three sets of four practice trials. Each set of practice trials included two congruent trials (one with all arrows pointing to the left and one with all arrows pointing to the right) and two incongruent trials (one with a left-facing central arrow and one with a right-facing central arrow). In order to pass the practice and progress to the test or scored trials, the child had to have three or more correct practice trials within a single set of four practice trials. If the child did not pass the first set of practice trials, a second set was presented. If the child did not pass the second set of practice trials, a third set of practice trials was administered. If the child was not able to pass any of the three sets of practice trials, the Flanker ended before any actual scored trials were presented and the child moved into the science assessment.

Before the practice trials started, children were presented with a screen providing the same standardized instructions that are described above for the test trials, which the assessor read. As noted above, the instructions stated, "Keep your eyes on the star. Answer as fast as you can without making mistakes. If you make a mistake, just keep going." The practice trials were like the subsequent test trials in

that a star appeared first on the screen to act as a focal point and a recorded female voice said "middle" to remind the child to look at and indicate the direction of the middle arrow. However, unlike in the test trials, during the practice trials the recorded voice was used to provide feedback to the child. If the child answered a practice trial correctly, the recorded voice said "That's right!" If the child did not respond correctly to a practice trial, the recorded voice provided feedback to the child to explain the correct answer and why it was correct.

Item-level data for the 20 scored test trials are included in the data file. Data are provided for four aspects of each test trial: (1) correct versus incorrect responses (C*FLKACC1-C*FLKACC20); (2) the type of trial, reported as congruent (more frequently presented but not included in reaction time scores; central arrow faces in the same direction as the flanking arrows) or incongruent (less frequently presented and used to calculate reaction time scores; central arrow faces in the direction opposite from the flanking arrows) (C*FLKCIC1-C*FLKCIC20); (3) reaction time reported in milliseconds (C*FLKRT1-C*FLKRT20); and (4) the direction that the central arrow faces (C*FLKARW1-C*FLKARW20).[13] Therefore, there are four variables associated with each of the 20 test trials. Children who did not pass any of the three sets of practice trials do not have item-level data because the item-level data correspond to the actual scored trials. Variable names for the item-level data begin with "C8" for spring fourth grade and "C9" for spring fifth grade.

The overall computed score reported for the fourth- and fifth-grade Flanker is derived using a formula provided by the task developer and follows the scoring algorithm used for this task in the NIH Toolbox (see *NIH Toolbox Scoring and Interpretation Guide* [Slotkin, Nowinski et al. 2012] for additional information on scoring). This is the same formula used to score the computerized DCCS score, adjusted for task parameters (number of administered trials). Like the DCCS, the overall Flanker score ranges from 0 to 10, with weight given to accuracy (0 to 5 units) and reaction time (0 to 5 units) in the computation of scores. Accuracy is considered first. If the child's accuracy rate is less than or equal to 80 percent, the child's overall computer score is based entirely on accuracy. If the child's accuracy rate is more than 80 percent, the child's overall computed score is based on a combination of accuracy and reaction time. Children who did not pass any of the three sets of practice trials do not have an overall Flanker score.

The accuracy score factored into the computation of the overall score can range from 0 to 5. Because all children used the Flanker start point for children 8 years and older, each child who successfully passed the practice was administered 20 test trials and was automatically given 20 accuracy points for 20 trials that are only administered to children younger than 8 years old. Therefore, there are a total of 40

[13] A variable to describe the direction that the central arrow faces is not necessary for analyzing task performance. It is included on the data file to allow researchers to reconstruct the exact trials that were presented in case there is interest in doing so.

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). The accuracy component of the overall computed Flanker score is calculated as follows:

Flanker accuracy score = 0.125 * number of correct responses.[14]

If the child's accuracy rate is higher than 80 percent, a reaction time score is added to the child's accuracy score.[15] Like the accuracy score, the reaction time score ranges from 0 to 5 points.

The reaction time component of the overall computed score for the Flanker is computed using the child's median reaction time to correct incongruent trials (i.e., the trials with the flanking arrows facing in a direction opposite the central arrow), following the same scoring algorithm outlined in the scoring manual for the NIH Toolbox (Slotkin, Nowinski, et al. 2012). First, for those children with greater than 80 percent accuracy on the 40 trials, the median reaction time is calculated based on reaction times for correct incongruent trials with reaction times greater than or equal to 100 milliseconds (msec) and within plus or minus three standard deviations from the child's mean reaction time on the correct incongruent 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 were 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 range of 0 to 5 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:

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

[14] The number of correct responses = 20 + the number of correct arrow trials out of the 20 administered trials. Thus, once the child has passed the practice trials and advanced into the scored portion of the assessment, 20 accuracy points are automatically awarded due to the chosen start point for the task. For this reason, it is not possible for ECLS-K:2011 children to get an accuracy score of 0. Therefore, the minimum possible value for the Flanker accuracy score is 2.5, and the maximum possible Flanker accuracy score is 5.

[15] The criterion of *greater than* 80 percent accuracy is calculated based on all 40 trials (20 administered arrow trials plus the 20 that are only administered to children younger than 8 years old). 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 20 administered arrow trials are required. If the criterion is 80 percent of the 40 trials, this translates to 12 of the 20 administered trials. For example, if a child responds accurately on 13 of the 20 administered arrow trials, the child's accuracy rate equals 82.5 percent (20 points automatically awarded for the nonadministered 20 trials plus the 13 correct arrow trials divided by 40; 33/40 = .825). In this example, the child's accuracy score would be [(20 + 13) * .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 Flanker. Alternatively, if the child responded accurately on 12 of the 20 administered arrow 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 = [(20 + 12) * .125] = 4).

where *RT* is the median reaction time on incongruent trials within set outer limits.[16]

To summarize, the overall computed score on the computerized Flanker 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 overall 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 incongruent 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).

It is important for researchers using the Flanker data to be aware of the characteristics of the overall Flanker 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 accuracy. The use of this scoring model with the data collected from children in the ECLS-K:2011 resulted in a non-normal distribution. There were 48 children in fourth grade and 32 children in fifth grade who failed to achieve greater than 80 percent accuracy (0.4 percent in fourth grade and 0.3 percent in fifth grade). The score for these children is calculated based solely on accuracy. There are 27 children in the fourth-grade data collection (0.2 percent) and 11 children in the fifth-grade data collection who met the accuracy threshold but did not have any correct incongruent trials; therefore, their score was set equal to their accuracy score because it was not possible to have a reaction time score for correct, incongruent trials. Thus, there were a total of 75 children (48 + 27) in fourth grade (0.6 percent) and 43 children (32 + 11) in fifth grade (0.4 percent) whose overall Flanker score is based on accuracy alone. The remaining children (99.4 percent in fourth grade and 99.6 percent in fifth grade) who have a computed overall score 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 Flanker 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. Users who want to analyze all children using the score indicating accuracy alone should recognize that this score is highly skewed, as most children were able to indicate the direction the central arrow was pointing 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.

[16] 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 incongruent 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 incongruent trials, the child's overall computed score on the Flanker was set equal to the child's accuracy score, and reaction time was not factored into the child's score.

The decision about how best to use the Flanker 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 flanker data, in deciding how best to utilize the ECLS-K:2011 Flanker data.

6.3.1 Mean Scores for the Flanker

This section presents the means for the Flanker scores in spring fourth and spring fifth grades. Means for the overall computed Flanker score by data collection round and by selected child characteristics are also presented.

Five scores based on the Flanker are presented on the data file for spring fourth grade and spring fifth grade: overall score (X*FLANKER; range: 0-10); accuracy score (X*FLKACC; range: 0-5) that is scaled as described above to compute the overall Flanker score; reaction time score (X*FLKICRT; range: 0-5) that is scaled to compute the overall Flanker score; count of correct, congruent trials (X*FLKCAC; range 0-13); and count of correct incongruent trials (X*FLKICAC; range 0-7). Researchers should note that the count of correct congruent trials and the count of correct incongruent trials represent accuracy by trial type for the 20 administered trials and are different from the total accuracy score (X*FLKACC, Flanker Accuracy Component [0-5] Scr) that is derived to compute the overall Flanker score. Researchers should also note that the reaction time score was only computed for cases for which the accuracy score was greater than 80 percent. If the accuracy score was not greater than 80 percent, then the reaction time score was set to -9 (not ascertained).

Table 6-9 presents the Flanker variable names, descriptions, value ranges, weighted means, and standard deviations for the spring of fourth grade and the spring of fifth grade.

Table 6-9. Flanker variable names, descriptions, value ranges, weighted means, and standard deviations for spring fourth grade and spring fifth grade: Spring 2015 and spring 2016

Variable name	Description	*n*	Range of possible values[1]	Weighted mean	Standard deviation
X8FLANKER	X8 Flanker Computed (Overall) Score	12,009	0–10	7.98	0.984
X8FLKACC	X8 Flanker Accuracy Component (0–5) Scr	12,009	0–5	4.96	0.129
X8FLKICRT	X8 Flanker Incon RT Component (0–5) Scr	11,934	0–5	3.03	0.923
X8FLKCAC	X8 Flanker Congruent Accuracy Count	12,009	0–13	12.93	0.484
X8FLKICAC	X8 Flanker Incongruent Accuracy Count	12,009	0–7	6.78	0.770
X9FLANKER	X9 Flanker Computed (Overall) Score	11,399	0–10	8.41	0.872
X9FLKACC	X9 Flanker Accuracy Component (0–5) Scr	11,399	0–5	4.97	0.107
X9FLKICRT	X9 Flanker Incon RT Component (0–5) Scr	11,355	0–5	3.45	0.830
X9FLKCAC	X9 Flanker Congruent Accuracy Count	11,399	0–13	12.94	0.389
X9FLKICAC	X9 Flanker Incongruent Accuracy Count	11,399	0–7	6.81	0.643

[1] Because 20 accuracy points are automatically awarded due to the chosen start point for the task, it is not possible for ECLS-K:2011 children to obtain an accuracy score of 0. Therefore, the lowest accuracy component (0-5) score in the data file is 2.5**,** and the lowest computed (overall) score in the data file is also 2.5.
NOTE: Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2015 and spring 2016.

Table 6-10 presents the mean overall Flanker scores for spring fourth and spring fifth grades by data collection round and by selected child characteristics.

Table 6-10. Mean Flanker overall score for spring fourth grade and spring fifth grade, by data collection round and child characteristics: Spring 2015 and spring 2016

Characteristic	Spring fourth grade (X8FLANKER)			Spring fifth grade (X9FLANKER)		
	n	Mean	*SD*	*n*	Mean	*SD*
Total sample	**12,009**	**7.98**	**0.984**	**11,399**	**8.41**	**0.872**
Sex						
Male	6,108	8.04	0.995	5,824	8.46	0.902
Female	5,886	7.92	0.968	5,562	8.36	0.835
Race/ethnicity						
White, non-Hispanic	5,683	8.07	0.913	5,408	8.50	0.788
Black, non-Hispanic	1,266	7.68	1.140	1,111	8.10	1.090
Hispanic	3,357	7.87	1.003	3,244	8.33	0.880
Asian, non-Hispanic	1,033	8.26	0.963	1,000	8.69	0.805
Hawaiian, Other Pacific Islander, non-Hispanic	66	8.27	0.830	62	8.72	0.592
American Indian/Alaska Native, non-Hispanic	103	8.18	0.746	101	8.49	0.722
Two or more races, non-Hispanic	492	8.09	0.990	464	8.52	0.850
School type						
Public school	10,893	7.97	0.992	10,335	8.40	0.876
Private school	1,057	8.12	0.881	1,006	8.52	0.819

NOTE: Spring fourth-grade estimates (X8) are weighted by W8C8P_20. Spring fifth-grade estimates (X9) are weighted by W9C9P_20. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. Detail may not sum to totals 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 2015 and spring 2016.

6.3.2 Use of Flanker Scores in Longitudinal Analyses

The Flanker was administered in the fourth- and fifth-grade data collections. The same task was used in both rounds and performance can be directly compared. Differences in scores indicate raw improvement or decline between rounds. A change in the child's score from one round to another represents real change in the level of performance for that child.

Like the computerized version of the DCCS, the total score (X*FLANKER) is a computed score that ranges from 0 to 10 and is based on both accuracy and reaction time. A score between 0 and 5 indicates that the child had lower accuracy (80 percent correct or less) and did not trade speed for accuracy (i.e., did not slow down as necessary in order to get the trial correct). This means that accuracy and not

reaction time is the better indicator of performance. A score between 5 and 10 indicates accuracy was high and reaction time was factored into the score. When a child is developmentally able to trade speed for accuracy (i.e., slow down to be accurate), then reaction time becomes the critical aspect of the scoring and is factored into the scoring equation.

The ECLS-K:2011 does not provide age-adjusted or grade-adjusted scale scores. This is left to the discretion of the data user, based on the research question.

6.3.3 Flanker Data Flags

There are two flags to indicate the presence or absence of Flanker data. X8FLNKFLG indicates the presence of data for the overall computed Flanker score (X8FLANKER) for the spring of fourth grade, and X9FLNKFLG indicates the presence of data for the overall computed Flanker score (X9FLANKER) for the spring of fifth grade. The flag values indicate whether the task was administered, whether the overall computed Flanker score is present and, if a score is not present, the reason why it is not present. Reasons why a score is not present when the Flanker was administered include failing the practice trials or having an administrative breakoff (meaning 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 (for example, a fire drill or the child needing to return to class) that interrupted an assessment session.

The Flanker flags for the spring of fourth grade and the spring of fifth grade have five possible values. A description of the values of the completion flags is provided in exhibit 6-3. The flag is equal to system missing when the child was not a participant in the round of data collection.

Exhibit 6-3. Data flag description for the Flanker for the spring of fourth grade and spring of fifth grade: Spring 2015 and spring 2016

X8FLNKFLG/X9FLNKFLG	Value
Not Administered	0
Flanker computed (overall) score present	1
Failed Arrows practice	2
Breakoff before passing practice trials	3
Breakoff after passing practice trials	4

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2015 and spring 2016.

6.4 Correlations Among Executive Function Measures

Table 6-11 shows across-round correlations by task. Correlations are provided across rounds for each task using the total score for the DCCS (X*DCCSTOT for the physical version of the DCCS and X*DCCSSCR for the computerized version of the task), the *W* score for Numbers Reversed (X*NRWABL), and the total score for the Flanker (X*FLANKER). For each task, note that correlations are highest near the diagonal, which is the correlation for the round immediately prior to a given round. The correlation, while it remains significant, gets progressively lower with the more time that passes. For example, the best predictor (i.e., highest correlation) of the spring fourth-grade (round 8) DCCS score is the spring third-grade (round 7) DCCS score, with a correlation coefficient of .52. Previous DCCS scores are also significantly correlated with the spring fourth-grade DCCS, but the relation becomes weaker going back in time. While DCCS performance in prior rounds significantly predicts DCCS performance in later rounds, other factors in intervening years presumably have an important influence as well. Measures of child, family and school characteristics that relate to performance on executive function are provided in the ECLS-K:2011 database. Exploration of the role these variables play in predicting executive function are beyond the scope of this report.

Table 6-11. Correlations of executive function scores across rounds, by task: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

EF Task	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9
DCCS									
Round 1: X1DCCSTOT	1.00	†	†	†	†	†	†	†	†
Round 2: X2DCCSTOT	.31	1.00	†	†	†	†	†	†	†
Round 3: X3DCCSTOT	.26	.32	1.00	†	†	†	†	†	†
Round 4: X4DCCSTOT	.26	.27	.33	1.00	†	†	†	†	†
Round 5: X5DCCSSCR	.24	.31	.28	.29	1.00	†	†	†	†
Round 6: X6DCCSSCR	.24	.25	.26	.28	.56	1.00	†	†	†
Round 7: X7DCCSSCR	.20	.23	.21	.25	.45	.49	1.00	†	†
Round 8: X8DCCSSCR	.18	.21	.21	.23	.40	.45	.52	1.00	†
Round 9: X9DCCSSCR	.17	.19	.19	.21	.36	.40	.47	.53	1.00
Numbers Reversed									
Round 1: X1NRWABL	1.00	†	†	†	†	†	†	†	†
Round 2: X2NRWABL	.57	1.00	†	†	†	†	†	†	†
Round 3: X3NRWABL	.50	.57	1.00	†	†	†	†	†	†
Round 4: X4NRWABL	.43	.52	.55	1.00	†	†	†	†	†
Round 5: X5NRWABL	.41	.47	.55	.58	1.00	†	†	†	†
Round 6: X6NRWABL	.39	.45	.49	.54	.58	1.00	†	†	†
Round 7: X7NRWABL	.39	.43	.46	.50	.54	.59	1.00	†	†
Round 8: X8NRWABL	.38	.41	.43	.48	.52	.56	.63	1.00	†
Round 9: X9NRWABL	.38	.41	.42	.47	.50	.54	.61	.66	1.00
Flanker									
Round 8: X8FLANKER	†	†	†	†	†	†	†	1.00	†
Round 9: X9FLANKER	†	†	†	†	†	†	†	.54	1.00

† Not applicable.
NOTE: All estimates unweighted. All correlations $p < .0001$. Data collection rounds are numbered as follows: Round 1 (fall kindergarten), round 2 (spring kindergarten), round 3 (fall first grade), round 4 (spring first grade), round 5 (fall second grade), round 6 (spring second grade), round 7 (spring third grade), round 8 (spring fourth grade), and round 9 (spring fifth grade). The physical version of the DCCS was administered in rounds 1-4, and the computerized version of the DCCS was administered in rounds 5-9. DCCS= Dimensional Change Card Sort.
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, spring 2013, spring 2014, spring 2015, and spring 2016.

Correlations across different executive function measures within rounds are presented in table 6-12. Correlations among the total score for the DCCS (X*DCCSTOT for the physical version of the DCCS and X*DCCSSCR for the computerized version of the task), the *W* score for Numbers Reversed (X*NRWABL), and the total score for the Flanker (X*FLANKER) are provided. The within-round correlations between DCCS and Numbers Reversed range from .30 to .40, with the highest correlations in second grade. The correlation between DCCS and Flanker ranges from .45 to .50. The correlation between Numbers Reversed and Flanker ranges from .24 to .25 These correlations show that while these measures are correlated, they do not entirely overlap and each accounts for unique aspects of executive function. For example, a task that primarily measures cognitive flexibility also requires the ability to focus attention to

relevant detail about the task and the ability to inhibit a quick response. Depending upon the research question, data users may be interested in using these executive function measures individually, or they may be interested in combining the three measures of executive function to create an overall measure of executive function, which may be more robust.

Table 6-12. Correlations of executive function scores across tasks, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Data collection round	DCCS and Numbers Reversed	DCCS and Flanker	Numbers Reversed and Flanker
Round 1: Fall kindergarten	.30	†	†
Round 2: Spring kindergarten	.30	†	†
Round 3: Fall first grade	.31	†	†
Round 4: Spring first grade	.31	†	†
Round 5: Fall second grade	.40	†	†
Round 6: Spring second grade	.36	†	†
Round 7: Spring third grade	.32	†	†
Round 8: Spring fourth grade	.32	.45	.24
Round 9: Spring fifth grade	.30	.50	.25

† Not applicable.
NOTE: All estimates unweighted. The total score for the physical version of the DCCS (X*DCCSTOT) was used in round 1 through round 4, and the total score for the computerized version of the DCCS (X*DCCSSCR) was used in correlations for round 5 through round 9. The variable used for Numbers Reversed was X*NRWABL for all rounds. The total score for the Flanker (X*FLANKER) was used in correlations. All correlations $p < .0001$.
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, spring 2013, spring 2014, spring 2015, and spring 2016.

6.5 Examination of Possible Assessor Effects in the Administration of the Executive Function Measures

In the ECLS-K:2011, assessors conducted one-on-one assessments with children. Providing individual attention to children was a strength of the study. Although one-on-one assessments are the preferred method of data collection with young children, they introduce a potential source of variance that is not related to the child's ability. Individual assessments can lead to sources of variance that are unique to individual assessors. Children's performance on tasks that are harder to administer and that rely more on clinical skill and judgment is more likely to be influenced by the skills of the assessor. Variation in assessment administration would be expected to be small for more automated or controlled assessments that were designed to minimize the role of the assessor such as the direct cognitive assessment, for example, by having the computer tell the assessor what to administer and having very standardized administration procedures. However, assessments for the executive function measures required more clinical skill and judgment and required the assessor to play a larger role in implementing key elements of the task. For

example, the assessor controlled the pace of the number presentation in Numbers Reversed. In the physical version of the DCCS administered in the kindergarten and first-grade rounds, the assessor had to correctly instruct the child, give the cards to the child in the correct order, make sure the child did not see the cards in advance, and record how the child sorted the cards. In the early rounds of the ECLS-K:2011, project staff observed variation in the administration of the executive function assessments by assessor. After these observations, there were changes in training of assessors and the administration of the assessments to make the executive function assessments more standardized. To explore whether assessors influenced performance on the executive function measures, assessor effects were examined in all rounds of the study, including those not covered by the current report. The analyses conducted are described below.

Each assessor (variable name F*CASSOR) for the ECLS-K:2011 conducted child assessments with multiple children within multiple schools and was assigned to a group of schools referred to as a work area (variable name F*CWKARE). Because of this design, the data have multiple levels, with students embedded within assessors (groups of students with the same assessor), students embedded within schools (groups of students in the same school), and assessors embedded within work areas (groups of assessors in the same work area). Cases within a group (e.g., students within the same school) may be more similar to each other than cases from different groups (e.g., students from different schools). Multilevel analyses take into account within-group similarity and do not assume that students or groups are independent, as traditional multiple regression analyses would.

A series of cross-sectional multilevel models were used to test possible assessor effects while partitioning out school and work area effects for three executive function measures in the ECLS-K:2011: the Numbers Reversed subtest of the *Woodcock-Johnson III* (WJ III) *Tests of Cognitive Abilities* (Woodcock, McGrew, and Mather 2001), the Dimensional Change Card Sort *(*DCCS*)* (Zelazo 2006; Zelazo et al. 2013), and the Flanker (Zelazo et al. 2013). The multilevel models used unweighted data and examined how much variability in the executive function outcomes was accounted for by assessors, measured by the significance of the variance and the magnitude of the Intra-Class Correlation Coefficient (ICC; readers may refer to Brunton-Smith, Sturgis, and Leckie 2017; Hox 1994; Raudenbush and Bryk 2002; and Singer and Willet 2003 for details about detecting assessor effects, ICC, and its utilities in research). The ICC of a two-level factor (e.g., assessor) represents the proportion of the outcome variance accounted for by assessors, which is defined and computed by (the variance estimate of a higher level factor[17] / total variance of the outcome)*100. This proportion is above and beyond the variance explained by the student-level covariates. Student-level covariates were included in all models.[18] An ICC of 2 percent to 3 percent or lower may be

[17] A higher level factor in this study may be level-2 assessor or level-3 work area.

[18] Covariates included in the analyses as fixed effects were the child's background characteristics (sex, age, and race), the child's math theta score, the highest education level of the parent(s), and household income. Random effects in two-level models were assessors. Random effects in three-level models were assessors and work area.

considered small, but the ICC is only one component of the assessor effect as the overall effect depends on other factors such as the average number of children assessed by an interviewer.

6.5.1 Numbers Reversed

First, two-level models with students embedded within assessors were run to examine assessor effects (level 1 = student, level 2 = assessor). In two-level models for Numbers Reversed scores with student covariates, the variance estimates for assessors were significant in all rounds; however, the ICCs for assessor effects were trivial (1.6 percent or less) (table 6-13). Estimates for assessors may have been significant due to large sample sizes.

Because assessors were embedded within work areas, three-level models were also tested to examine assessors and work areas in the same model. In three-level models (level 1 = student, level 2 = assessor, level 3 = work area) with student-level covariates, although assessor effects were significant in most rounds, the ICCs for assessor effects were also trivial (1.5 percent or less) (table 6-14). A three-level model with assessor and school was not run because assessors worked in more than one school and thus were not embedded within particular schools as they were within work areas.

Table 6-13. Variance estimates in two-level models for Numbers Reversed *W*-ability scores and student-level covariates, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Data collection round	Number of students[1]	Number of assessors	Assessor effects		
			Level 1 (student)	Level 2 (assessor)	ICC percent (assessor)
Round 1	11,185	292	545.53***	14.53***	1.6
Round 2	13,046	284	530.66***	14.10***	1.5
Round 3	4,051	98	507.10***	11.95***	1.4
Round 4	12,429	272	418.63***	5.96***	0.9
Round 5	3,947	109	377.11***	4.35*	0.7
Round 6	10,567	271	354.65***	5.99***	1.1
Round 7	10,399	274	343.61***	2.52**	0.5
Round 8	9,867	275	324.47***	3.26**	0.7
Round 9	9,325	269	340.36***	6.05***	1.2

*$p < .05$. **$p < .01$. ***$p < .001$

[1] The number of students shown in the table reflects those with data for variables in the model and are less than the total number of students in a round. In addition, the numbers of students in rounds 3 and 5 are less than the numbers of students in other rounds because these rounds were conducted with a subsample of children.

NOTE: The variance estimates do not sum to the total outcome variance because the variance estimates have controlled for student-level covariates (that is, the part of outcome variance explained by student-level covariates are already excluded from these variance estimates). Covariates included in the analyses were the child's background characteristics (sex, age, and race), the child's math theta score, the highest education level of the parent(s), and household income. ICC = Intra-Class Correlation. The adjusted ICC was computed by (level 2 variance estimate / total variance of the outcome)*100, where the model-based total outcome variance is the sum of the variance estimates from the two-level model of an outcome without covariates. Data collection rounds are numbered as follows: Round 1 (fall kindergarten), round 2 (spring kindergarten), round 3 (fall first grade), round 4 (spring first grade), round 5 (fall second grade), round 6 (spring second grade), round 7 (spring third grade), round 8 (spring fourth grade), and round 9 (spring fifth grade).

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, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 6-14. Variance estimates in three-level models for Numbers Reversed *W*-ability scores and student-level covariates, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Data collection round	Number of students[1]	Number of assessors	Number of work areas	Level 1 (student)	Level 2 (assessor)	ICC percent (assessor)	Level 3 (work area)	ICC percent (work area)
Round 1	11,185	292	88	544.83***	13.84***	1.5	1.51	0.2
Round 2	13,046	284	90	529.69***	11.43***	1.2	3.80*	0.4
Round 3	4,051	98	31	505.97***	9.07**	1.1	4.39	0.5
Round 4	12,429	272	89	418.75***	5.87***	0.9	0.00	0.0
Round 5	3,947	109	34	377.44***	3.98*	0.7	0.00	0.0
Round 6	10,567	271	89	354.66***	5.20***	1.0	0.79	0.1
Round 7	10,399	274	89	343.73***	1.50	0.3	0.91	0.2
Round 8	9,867	275	89	324.65***	2.31*	0.5	0.76	0.2
Round 9	9,325	269	89	341.06	1.52	0.3	3.87***	0.8

*$p < .05$. **$p < .01$. ***$p < .001$

[1] The number of students shown in the table reflects those with data for variables in the model and are less than the total number of students in a round. In addition, the numbers of students in rounds 3 and 5 are less than the numbers of students in other rounds because these rounds were conducted with a subsample of children.

NOTE: The variance estimates do not sum to the total outcome variance because the variance estimates have controlled for student-level covariates (that is, the part of outcome variance explained by student-level covariates are already excluded from these variance estimates). Covariates included in the analyses were the child's background characteristics (sex, age, and race), the child's mathematics theta score, the highest education level of the parent(s), and household income. ICC = Intra-Class Correlation. The adjusted ICC was computed by (level 2 variance estimate / total variance of the outcome) * 100, where the model-based total outcome variance is the sum of the variance estimates from the three-level model of an outcome without covariates. Data collection rounds are numbered as follows: Round 1 (fall kindergarten), round 2 (spring kindergarten), round 3 (fall first grade), round 4 (spring first grade), round 5 (fall second grade), round 6 (spring second grade), round 7 (spring third grade), round 8 (spring fourth grade), and round 9 (spring fifth grade).

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, spring 2013, spring 2014, spring 2015, and spring 2016.

6.5.2 DCCS

For the DCCS, total scores were examined for rounds 1 through 4 (X*DCCSTOT)[19] and total overall scores were used for later rounds (X*DCCSSCR). In two-level models (level 1 = student, level 2 = assessor) for DCCS scores with student covariates, estimates for assessors were significant, but the ICCs for assessor effects were trivial (2 percent or less) in most rounds except rounds 1, 2, and 4 (range of 3.8 to 7.5 percent) (table 6-15). In three-level models (level 1 = student, level 2 = assessor, level 3 = work area) for DCCS scores with student covariates, estimates for assessors also were significant, but the ICCs for assessor effects were trivial (1.6 percent or less) in most rounds except rounds 1, 2, and 4 (range of 3.2 to 7.7 percent) (table 6-16). The greater assessor effects in the early rounds of the study compared to later rounds may be explained by a change in DCCS administration. A table-top version of the DCCS was used in rounds 1 through 4, whereas a computerized version of the DCCS was used in later rounds. It is possible that greater assessor effects were present for the table-top version because it required more assessor input

[19] For more information about the DCCS total scores (X*DCCSTOT), see the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b). U.S. Department of Education. Washington, DC:National Center for Education Statistics.

than the computerized version used when children were older. Assessor effects likely decreased in some of the three-level models when work area was added because of a part of the variance shared by assessor effects and work area. However, work area effects were not significant in some models, and this can be interpreted to mean that assessors' assessment-related skills or actions within the same work area were not more similar than across work areas.

Taken together, results showed that assessor effects accounted for a small, but not meaningfully significant, proportion of the variance for Numbers Reversed in all rounds of the study. For the DCCS measures, assessor effects were small in most rounds, but larger in rounds 1, 2, and 4. In general, users do not need to address assessor effects in their analyses; however, users could consider controlling for them (variable name F*CASSOR) with the DCCS in early rounds of the study (rounds 1 and 2, and possibly 4). Further, consistent with our expectations, compared across two-level and three-level models, results suggest that, although ICCs of assessor effects were not large, they were generally still larger than ICCs of work area effects. That is, work area effects were smaller factors in explaining student score differences.

Table 6-15. Variance estimates in two-level models for DCCS scores and student-level covariates, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Data collection round	Number of students[1]	Number of assessors	Assessor effects		
			Level 1 (student)	Level 2 (assessor)	ICC percent (assessor)
Round 1	11,189	292	7.92***	0.83***	7.5
Round 2	13,047	284	5.84***	0.56***	7.2
Round 3	4,051	98	4.67***	0.12**	2.0
Round 4	12,429	272	4.08***	0.20***	3.8
Round 5	3,933	109	1.53***	0.01	0.3
Round 6	10,526	271	1.30***	0.03***	1.8
Round 7	10,301	274	1.69***	0.02***	1.1
Round 8	9,813	275	0.76***	0.01***	0.9
Round 9	9,289	269	0.74***	0.01***	1.1

*$p < .05$. **$p < .01$. ***$p < .001$

[1] The number of students shown in the table reflects those with data for variables in the model and are less than the total number of students in a round. In addition, the numbers of students in rounds 3 and 5 are less than the numbers of students in other rounds because these rounds were conducted with a subsample of children.

NOTE: The variance estimates do not sum to the total outcome variance because the variance estimates have controlled for student-level covariates (that is, the part of outcome variance explained by student-level covariates are already excluded from these variance estimates). Covariates included in the analyses were the child's background characteristics (sex, age, and race), the child's math theta score, the highest education level of the parent(s), and household income. ICC = Intra-Class Correlation. The adjusted ICC was computed by (level 2 variance estimate / total variance of the outcome) * 100, where the model-based total outcome variance is the sum of the variance estimates from the two-level model of an outcome without covariates. Data collection rounds are numbered as follows: Round 1 (fall kindergarten), round 2 (spring kindergarten), round 3 (fall first grade), round 4 (spring first grade), round 5 (fall second grade), round 6 (spring second grade), round 7 (spring third grade), round 8 (spring fourth grade), and round 9 (spring fifth grade).

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, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 6-16. Variance estimates in three-level models for DCCS and student-level covariates, by data collection round: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Data collection round	Number of students[1]	Number of assessors	Number of work areas	Level 1 (student)	Level 2 (assessor)	ICC percent (assessor)	Level 3 (work area)	ICC percent (work area)
Round 1	11,189	292	88	7.90***	0.84***	7.5	0.00	0.0
Round 2	13,047	284	90	5.82***	0.61*	7.7	0.00	0.0
Round 3	4,051	98	31	4.66***	0.10**	1.6	0.03	0.5
Round 4	12,429	272	89	4.08***	0.17***	3.2	0.05*	0.9
Round 5	3,933	109	34	1.53***	0.00	0.1	0.00	0.2
Round 6	10,526	271	89	1.30***	0.01*	0.6	0.02***	1.3
Round 7	10,301	274	89	0.96***	0.01	0.6	0.01**	0.5
Round 8	9,813	275	89	0.77***	0.00	0.2	0.01**	0.7
Round 9	9,289	269	89	0.74***	0.00	0.3	0.01**	0.8

*$p < .05$. **$p < .01$. ***$p < .001$

[1] The number of students shown in the table reflects those with data for variables in the model and are less than the total number of students in a round. In addition, the numbers of students in rounds 3 and 5 are less than the numbers of students in other rounds because these rounds were conducted with a subsample of children.

NOTE: The variance estimates do not sum to the total outcome variance because the variance estimates have controlled for student-level covariates (that is, the part of outcome variance explained by student-level covariates are already excluded from these variance estimates). Covariates included in the analyses were the child's background characteristics (sex, age, and race), the child's math theta score, the highest education level of the parent(s), and household income. ICC = Intra-Class Correlation. The adjusted ICC was computed by (level 2 variance estimate / total variance of the outcome)*100, where the model-based total outcome variance is the sum of the variance estimates from the three-level model of an outcome without covariates. Data collection rounds are numbered as follows: Round 1 (fall kindergarten), round 2 (spring kindergarten), round 3 (fall first grade), round 4 (spring first grade), round 5 (fall second grade), round 6 (spring second grade), round 7 (spring third grade), round 8 (spring fourth grade), and round 9 (spring fifth grade).

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, spring 2013, spring 2014, spring 2015, and spring 2016.

6.5.3 Flanker

The Flanker was administered in spring fourth grade (round 8) and spring fifth grade (round 9) of the study. Two-level models with students embedded within assessors were run to examine assessor effects (level 1 = student, level 2 = assessor). In two-level models for Flanker scores with student covariates, the variance estimates for assessors were significant in both rounds; however, the ICCs for assessor effects were trivial (1.7 percent) (table 6-17). Estimates for assessors may have been significant due to large sample sizes.

Three-level models were also tested to examine assessors and work areas in the same model. In three-level models (level 1 = student, level 2 = assessor, level 3 = work area) with student-level covariates, although assessor effects were significant in both rounds 8 and 9, the ICCs for assessor effects were also trivial (0.7 percent or less) (table 6-18). A three-level model with assessor and school was not run because assessors worked in more than one school and thus were not embedded within particular schools as they were within work areas.

Table 6-17. Variance estimates in two-level models for Flanker scores and student-level covariates, by data collection round: Spring 2015 and spring 2016

Data collection round	Number of students[1]	Number of assessors	Assessor effects		
			Level 1 (student)	Level 2 (assessor)	ICC percent (assessor)
Round 8	9,810	275	0.83***	0.02***	1.7
Round 9	9,301	269	0.63***	0.01***	1.7

*$p < .05$. **$p < .01$. ***$p < .001$

[1] The number of students shown in the table reflects those with data for variables in the model and are less than the total number of students in a round.

NOTE: The variance estimates do not sum to the total outcome variance because the variance estimates have controlled for student-level covariates (that is, the part of outcome variance explained by student-level covariates are already excluded from these variance estimates). Covariates included in the analyses were the child's background characteristics (sex, age, and race), the child's math theta score, the highest education level of the parent(s), and household income. ICC = Intra-Class Correlation. The adjusted ICC was computed by (level 2 variance estimate / total variance of the outcome) * 100, where the model-based total outcome variance is the sum of the variance estimates from the two-level model of an outcome without covariates. Data collection rounds are numbered as follows: Round 8 (spring fourth grade) and round 9 (spring fifth grade).

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2015 and spring 2016.

Table 6-18. Variance estimates in three-level model for Flanker scores and student-level covariates, by data collection round: Spring 2015 and spring 2016

Data collection round	Number of students[1]	Number of assessors	Number of work areas	Level 1 (student)	Level 2 (assessor)	ICC percent (assessor)	Level 3 (work area)	ICC percent (work area)
Round 8	9,810	275	89	0.83***	0.01*	0.6	0.01***	1.2
Round 9	9,301	269	89	0.63***	0.01*	0.7	0.01***	1.1

*$p < .05$. **$p < .01$. ***$p < .001$

[1] The number of students shown in the table reflects those with data for variables in the model and are less than the total number of students in a round.

NOTE: The variance estimates do not sum to the total outcome variance because the variance estimates have controlled for student-level covariates (that is, the part of outcome variance explained by student-level covariates are already excluded from these variance estimates). Covariates included in the analyses were the child's background characteristics (sex, age, and race), the child's math theta score, the highest education level of the parent(s), and household income. ICC = Intra-Class Correlation. The adjusted ICC was computed by (level 2 variance estimate / total variance of the outcome)*100, where the model-based total outcome variance is the sum of the variance estimates from the three-level model of an outcome without covariates. Data collection rounds are numbered as follows: Round 8 (spring fourth grade) and round 9 (spring fifth grade).

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2015 and spring 2016.

6.5.4 Skew

Users should be aware that the executive function measures show some skew, which could have affected the analyses on assessor effects that were conducted. Skew was relatively small, and Numbers Reversed measures were less skewed than DCCS and Flanker measures (table 6-19). In rounds 1 through

4, an ordinal variable, X*DCCSCAT, created from X*DCCSTOT, was tested to reduce skew.[20] Although the X*DCCSCAT variables were less skewed, the results for assessor effects were similar. Thus, results for X*DCCSTOT are presented below. Although there was some variation in skew by measure and round, the findings for assessor effects were generally consistent across rounds.

Table 6-19. Skew of Numbers Reversed, DCCS, and Flanker scores

Variable	Skew
X1NRWABL	0.3948
X2NRWABL	-0.3055
X3NRWABL	-0.5067
X4NRWABL	-0.8627
X5NRWABL	-0.8859
X6NRWABL	-0.8989
X7NRWABL	-0.6647
X8NRWABL	-0.5236
X9NRWABL	-0.4077
X1DCCSTOT	-1.6082
X1DCCSCAT	-0.7051
X2DCCSTOT	-1.9874
X2DCCSCAT	-0.8771
X3DCCSTOT	-2.2219
X3DCCSCAT	-0.8871
X4DCCSTOT	-2.4435
X4DCCSCAT	-1.2497
X5DCCSSCR	-1.2615
X6DCCSSCR	-1.4469
X7DCCSSCR	-1.6618
X8DCCSSCR	-1.6559
X9DCCSSCR	-1.4928
X8FLANKER	-1.2071
X9FLANKER	-1.5081

NOTE: DCCS = Dimensional Change Card Sort
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, spring 2013, spring 2014, spring 2015, and spring 2016.

[20] In order to reduce the skew of DCCS total scores in rounds 1 through 4, categorical versions of the DCCS total scores were tested. Results were similar with categorical variables; thus only total scores are reported. These categorical versions were not appropriate for later rounds. The categorical measure in rounds 1-4, X*DCCSCAT, created from X*DCCSTOT, was constructed as follows:

- If (X*CSPRES = 0, 1, 2, 3, 4) then X*DCCSCAT= 0 (failed preswitch).
- If (X*CSPRES = 5 or 6) and (X*CSPSSC = 0, 1, 2, 3,4) then X*DCCSCAT= 1 (passed preswitch, failed postswitch).
- If (X*CSPRES = 5 or 6) and (X*CSPSSC = 5 or 6) and (X*CSBGSC = 0, 1, 2, 3,4) then X*DCCSCAT = 2 (passed preswitch, passed postswitch, failed border).
- If (X*CSPRES = 5 or 6) and (X*CSPSSC = 5 or 6) and (X*CSBGSC = 5 or 6) then X*DCCSCAT = 3 (passed preswitch, passed postswitch, passed border).

7. CHILD QUESTIONNAIRE

In the spring of third, fourth, and fifth grades, assessors asked children to complete a self-administered child questionnaire. Unlike the hard-copy child questionnaires that were administered during the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) by assessors who read the questions/items to the children, the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) child questionnaire was administered on a computer using audio computer-assisted self-interview (audio-CASI) technology and headphones. This chapter provides information about the administration of the questionnaire and the content for third, fourth, and fifth grades. Additional information about the design and administration of the child questionnaire can be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Third Grade Data File and Electronic Codebook, Public Version* (NCES 2018-034) (Tourangeau et al. 2018a); the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Fourth Grade Data File and Electronic Codebook, Public Version* (NCES 2018-032) (Tourangeau et al. 2018b); and the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Fifth Grade Data File and Electronic Codebook, Public Version* (NCES 2019-051) (Tourangeau et al. 2019).

After assessors assisted children with the headphones and the volume level, children listened as the software system read the instructions and questionnaire items. One questionnaire item at a time was displayed on the laptop's screen. In third grade a recorded human voice read each question and the response options to the child. In fourth and fifth grades, a computer-generated voice was used. The child responded by selecting the desired response on the laptop's touch screen. The audio-CASI questionnaire standardized administration and accommodated the variation in children's reading ability levels. It also allowed the child privacy to respond to the questions and limited distractions because the headphones worn during the administration minimized extraneous noise.

The child questionnaires administered in third grade, fourth grade, and fifth grade included a subset of items administered in the previous cohort study, the ECLS-K, as well as new content. Some constructs were administered across all grades, some constructs were included in one round of data collection, and some were included in two or more rounds of data collection. In some cases, the response options were modified from the original version of the question to simplify the questionnaire for students by reducing the number of variations in the response scales. In third grade, the child questionnaire had 37 questions and took approximately 11 minutes to complete. In fourth grade, the child questionnaire had 35

questions and took approximately 8 minutes to complete. In fifth grade, the child questionnaire had 48 items and took approximately 10 minutes to complete.

Exhibit 7-1 shows the content areas included in the third-grade, fourth-grade, and fifth-grade child questionnaires and the corresponding item-level variables along with their sources. Variable names for the item-level data begin with "C7" for spring third grade, "C8" for spring fourth grade, and "C9" for spring fifth grade. Many of the items in the child questionnaire were adapted from existing scales and were used with the permission of the author or publisher. Data for the individual items are included in the kindergarten through fifth-grade data file, but composite variables for each construct are not provided; it is left to analysts to decide how best to use these data in their analyses.

Although composites are not provided on the data file, some preliminary analyses were conducted to examine the internal consistency of the items for each construct. Each construct assessed in the child questionnaires is discussed and the internal consistency of the variables measuring each construct is provided when appropriate in the sections that follow.

Exhibit 7-1. Child questionnaire content and item-level variables for spring third grade, spring fourth grade, and spring fifth grade: Spring 2014, spring 2015, and spring 2016

Child questionnaire content area	Grade	Number of items	Item-level variable names
Perceived Interest/Competence in Reading[1]	3	5	C7LKREAD, C7INTREAD, C7CTWREAD, C7GDREAD, C7ENJREAD
Perceived Interest/Competence in Math[1]	3	5	C7LIKMTH, C7INTMTH, C7CTWMTH, C7GDMTH, C7ENJMTH
Perceived Interest/Competence in Science[1]	3	5	C7LKSCI, C7INTSCI, C7CTWSCI, C7GDSCI, C7ENJSCI
Peer Relationships[1]	3	6	C7HASFRNDS, C7MKFRNDS, C7GETALNG, C7EASYLIK, C7WTMEFRND, C7MORFRND
Peer Victimization[2]	3, 4, 5	4	C*TEASED, C*LIESABT, C*PUSHCH, C*EXCLDCH
Social Anxiety/Fear of Negative Evaluation[3]	3, 4, 5	3	C*WRYTHK, C*WRYDTLK, C*AFRDNTLK
Prosocial Behavior[4]	3	3	C7CHEERUP, C7HLPOTH, C7NICEOTH
Life Satisfaction[5]	3, 5	6, 3[6]	C7HAPHOB, C*HAPTHGS, C*HAPATTN, C7HAPFRND, C7HAPSKIL, C*HAPNBHD
Behavioral Engagement[7]	4, 5	5	C*TRYHRD, C*WRKHRD, C*PARDIS, C*PAYATT, C*LSTNCL
Peer Social Support[8]	4, 5	6	C*KIDBTR, C*KIDPLY, C*KIDHAP, C*KIDHLP, C*FRIEND, C*HELPMN
Loneliness[9]	4, 5	3	C*LONELY, C*LFTOUT, C*ALONE
Media Usage[10]	4, 5	3, 5[11]	C8OFTTXT/C9OFTTXT, C8RULWHO/C9RULWHO, C8RULWHN/C9RULWHN, C9ONLINE, C9SOCLNET
Pets[12]	4	18	C8CURPET, C8EVRPET, C8AGEPET, C8NUMPET, C8PETDOG, C8PETCAT, C8PETRAB, C8PETBRD, C8PETFSH, C8PETSNK, C8PETHRS, C8PETOTH, C8HVFVPET, C8FAVPET, C8PLYPET, C8PETHMW, C8PETSAD, C8PETFAM
School Belonging[13]	5	5	C9FITIN, C9CLOSCL, C9CLOSTC, C9ENJOY, C9SAFE
Grit[14]	5	6	C9FINISH, C9TRYMST, C9WKGOAL, C9WKHDQT, C9WKSETDO, C9TRYIMPRV
Worry/Stress About School[15]	5	5	C9WRYTST, C9HARDFIN, C9ASHAME, C9WRYWEL, C9WRYFIN
Parental Monitoring[16,17]	5	3	C9KNWFREE, C9KNWHW, C9KNWGRD

[1] Adapted from the Self Description Questionnaire I (SDQI) © Herbert Marsh. SELF Research Centre (Bankstown Campus), University of Western Sydney, Australia. Used with permission.

[2] Peer victimization items were adapted from a 21-item scale by Espelage, D.L., and Holt, M. (2001). Bullying and peer victimization during early adolescence: Peer influences and psychosocial correlates. *Journal of Emotional Abuse, 2*: 123–142.

[3] Adapted from the Social Anxiety Scale for Children—Revised ©1993 Annette M. La Greca, University of Miami. Used with permission. La Greca, A.M. and Stone, W.L. (1993). Social anxiety scale for children—revised: Factor structure and concurrent validity. *Journal of Clinical Child Psychology*, *22*(1): 17–27.

[4] Adapted from the Children's Social Behavior Scale—Self Report (CSBS-S). Crick, N.R., and Grotpeter, J.K. (1995). Relational aggression, gender, and social psychological adjustment. *Child Development, 66*: 710–722.
[5] Adapted from the National Institutes of Health (NIH) Toolbox for the Assessment of Neurological and Behavioral Function (version 1.0): Domain-Specific Life Satisfaction Survey from the NIH Toolbox Emotion Battery © 2012 Northwestern University and the National Institutes of Health. Used with permission.
[6] There were six items from the Domain-Specific Life Satisfaction Scale administered in third grade, but only a subset of those items was asked in fifth grade. Three of the six items were repeated in fifth grade.
[7] Adapted from Skinner, E.A., Kindermann, T.A., and Furrer, C.J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement*, *69*(3): 493-525.
[8] Adapted from Vandell, D. (2000). *Peer Social Support, Bullying, and Victimization* (Form FLV05G3: *Kids in My Class at School*) [measurement instrument]. National Institute of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development: Phase III, 2000-2004.
[9] Adapted from Parker, J.G., and Asher, S.R. (1993). Friendship and friendship quality in middle childhood: Links with peer group acceptance and feelings of loneliness and social dissatisfaction. *Developmental Psychology*, *29*(4): 611-621.
[10] Adapted from the *PEW September Tracking Survey 2009*. Princeton Survey Research Associates International. (2009). *PEW September Tracking Survey 2009*. Pew Internet and American Life Project.
[11] There were three items on media usage in fourth grade that asked children about frequency of online activity and family rules. These items along with two additional items about particular types of online activities were asked in fifth grade.
[12] Adapted from the CENSHARE Pet Attachment Survey. Holcomb, R., Williams, R.C., and Richards, P.S. (1985). The elements of attachment: Relationship maintenance and intimacy. *Journal of the Delta Society*, *2*(1): 28-34.
[13] Grade 8 Student Questionnaire, ECLS-K.
[14] Adapted from the Short Grit Scale in collaboration with Angela Duckworth for the ECLS-K:2011. Duckworth, A.L., and Quinn, P.D. (2009). Development and Validation of the Short Grit Scale (Grit–S). *Journal of Personality Assessment, 91*:2, 166-174.
[15] Adapted from the Internalizing Problems Scale that was developed for the ECLS-K and used in the ECLS-K grade 3 and grade 5 child-reported Self Description Questionnaire and the Grade 8 Student Questionnaire.
[16] Adapted from the Self-Disclosure & Parental Monitoring/Knowledge Scale (Kerr and Stattin, 2000). Kerr, M., and Stattin, H. (2000). What parents know, how they know it, and several forms of adolescent adjustment: Further support for a reinterpretation of monitoring. *Developmental Psychology, 36:* 366-380.
[17] In the spring of fourth grade, parents were also asked about parental monitoring of media usage. Parents were asked if they monitor how many hours their child spends online (P8MONTIM) and if they monitor what their child looks at online or what websites and accounts their child can join online (P8MONCON). These questions complement questions asked of the child on the child questionnaire.
NOTE: An asterisk "*" is a placeholder for the round number in variable names. Third grade is round 7, fourth grade is round 8, and fifth grade is round 9. For example, the variable C*TEASED is listed in the table; this indicates that the variables C7TEASED, C8TEASED, and C9TEASED are available in the dataset.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014, spring 2015, and spring 2016.

7.1 Self-Description Questionnaire (SDQ)

The child questionnaire administered in the third-grade data collection included items from the Self Description Questionnaire I (SDQI) (Marsh 1992) to assess both academic and nonacademic aspects of self-concept. Children rated their perceived interest and competence in reading, mathematics, and science (academic self-concept) and their peer relationships (nonacademic self-concept). The items administered in ECLS-K:2011 included a subset of items administered in ECLS-K, the prior cohort study, and also a modified set of items that asked about science rather than the set of items that asked about "all subjects" in the ECLS-K.

Children rated their perceived interest and competence in reading, mathematics, and science with five parallel questions for each subject. For example, the reading items were "I like READING," "I am interested in READING," "I cannot wait to READ each day,"[1] "I am good at READING," and "I enjoy doing work in READING." The mathematics and science items were the same, substituting "math" or "science" for "reading."

Children rated their peer relationships with six items that were selected verbatim from the original 9-item scale. The peer relationship items captured children's popularity with peers, how easily they make friends, and whether others want them as a friend. The ECLS-K:2011 third-grade child questionnaire included these items: "I have lots of friends," "I make friends easily," "I get along with kids easily," "I am easy to like," Other kids want me to be their friend," and "I have more friends than most other kids."

The reading, mathematics, science, and peer relationship items were adapted and used with permission. Items in the original SDQI were rated on a 5-point scale (False, Mostly false, Sometimes false/Sometimes true, Mostly true, True). In contrast, the ECLS-K:2011 items were rated on a 4-point scale (Not at all true, A little bit true, Mostly true, Very true). In addition, the reading, mathematics, science, and peer relationship items were presented in an order different from the original scale and an order different from the ECLS-K. The four sets of items were intermixed so that items on the same topic (i.e., reading, mathematics, science, and peer relationships) were not all asked together. The SDQI items were presented at the beginning of the child questionnaire in third grade.

Internal consistency reliability (Cronbach's alpha) was examined for each of the subscales in third grade. The internal consistency was .85 for perceived interest and competence in reading, .89 for

[1] The original wording of this item from the SDQ1 was "I look forward to READING."

perceived interest and competence in math, .86 for perceived interest and competence in science, and .78 for peer relationships.

7.2 Peer Victimization

In third grade, fourth grade, and fifth grade, children were asked four questions about peer victimization. These items were adapted from a 21-item scale by Espelage and Holt (2001) for the purpose of this study. The same questions were used in each round, and they were parallel to the questions asked of the teachers and parents. Teachers were asked the same types of peer victimization questions in the spring of second, third, fourth, and fifth grades,[2] and parents were asked these questions in second, third, and fourth grades.[3] Each of the four questions asked about a different type of peer victimization, and children were asked to respond how often they had experienced that type of victimization during the school year (Never, Rarely, Sometimes, Often, Very often). Children were asked the following questions: "During this school year, how often have other students teased you, made fun of you, or called you names?," "…how often have other students told lies or untrue stories about you?," "…how often have other students pushed, shoved, slapped, hit, or kicked you?," and "…how often have other students left you out from playing with them on purpose?"

These questions ask about specific types of victimization that can be considered separately or in combination depending on the research question. Whether these items are combined or not is left to the discretion of the data user. However, preliminary analyses were conducted to examine the internal consistency reliability of these items (Cronbach's alpha). The internal consistency of these four items is .74, .79, and .81, respectively, in the third-, fourth-, and fifth-grade rounds of data collection.

7.3 Social Anxiety/Fear of Negative Evaluation

Children were asked three items that were adapted from the Social Anxiety Scale for Children –Revised (La Greca and Stone 1993) in third, fourth, and fifth grades. These items assess Fear of Negative Evaluation, an aspect of social anxiety. Children were asked to respond using a 5-point scale

[2] In the spring of second, third, fourth, and fifth grades, teachers responded to an additional set of questions about peer victimization. In addition to asking teachers about the child as the victim of peer victimization, teachers were asked about their perceptions of the same types of peer victimization but with the child as the aggressor. See chapter 8 for a description of the teacher items administered.

[3] In the spring of second grade, parents were asked about only three of the four types of peer victimization. See chapter 8 for a description of the parent items administered.

(Never, Rarely, Sometimes, Often, Very often), which was an adaptation from the original scale that used different anchor points (Not at all, Hardly ever, Sometimes, Most of the time, All the time).

Internal consistency reliability (Cronbach's alpha) was examined for these three items. The internal consistency was .83 in third grade, .84 in fourth grade, and .88 in fifth grade.

7.4 Prosocial Behavior

In third grade, children were asked three items adapted from the Children's Social Behavior Scale–Self-Report (Crick and Grotpeter 1995) to assess children's prosocial behavior. Children were asked to respond on a 5-point scale (Never, Rarely, Sometimes, Often, Very often) to the following questions: "I try to cheer up other classmates who are upset or sad about something," "I help out other kids when they need it," and "I say or do nice things for other classmates." Within the third-grade child questionnaire, these items followed the peer victimization items and were intermixed with the fear of negative evaluation items. Internal consistency reliability for these items was examined using Cronbach's alpha. The items had an internal consistency of .76.

7.5 Life Satisfaction

Six items adapted from the National Institutes of Health (NIH) Toolbox Domain-Specific Life Satisfaction Survey were administered in the third grade. In fifth grade, three of the six items were administered. Children were asked to rate "how happy"[4] they were with different aspects of their life using a 5-point scale (Not at all, A little bit, Somewhat, Quite a bit, Very much). In third grade, children were asked about how happy they were with "hobbies and free time," "clothes, games, toys, and other things I have," "how much attention I get from my parents," "the friends I have," "my skills and talents," and "the neighborhood and community in which I live." In fifth grade, children were asked about how happy they were with "clothes, games, toys, and other things I have," "how much attention I get from my parents," and "the neighborhood and community in which I live." These items were included at the end of the third-grade questionnaire, following the parental monitoring items. The NIH Toolbox Domain-Specific Life Satisfaction Survey is a supplemental measure in the NIH Toolbox for the Assessment of Neurological and Behavioral Function (Slotkin, Nowinski et al. 2012). These items are not expected to

[4] The ECLS-K:2011 changed the wording of the NIH Toolbox items from "I am satisfied with…" to "I am happy with" to make the items more age appropriate. For example, children in ECLS-K:2011 were asked to rate the following statement, "I am happy with my hobbies and free time activities."

scale together (Slotkin, Nowinski et al. 2012) and therefore no internal consistency reliability information is presented.

7.6 Behavioral Engagement

In the fourth- and fifth-grade data collections, the child questionnaire included a 5-item scale to measure the child's behavioral engagement in the classroom (Skinner, Kindermann, and Furrer 2009). Children were asked to report their engagement using items that tapped their effort, attention, and persistence in learning activities. Children were asked to "Think about yourself and your experiences this school year. How often do the following things happen?" Children used a 5-point scale (Never, Rarely, Sometimes, Often, Very often) to rate the following items: "I try hard to do well in school," "In class, I work as hard as I can," "When I'm in class, I participate in class discussions," "I pay attention in class," and "When I'm in class, I listen carefully." These items in ECLS-K:2011 used the original wording reported in Skinner, Kindermann, and Furrer (2009), but the response options were adapted to match the response options for other items in the child questionnaire. The original response options ranged from "Not at all true" to "Very true." These five questions were asked at the beginning of the child questionnaire during the fourth- and fifth-grade data collections, immediately following the practice items.

Internal consistency reliability was examined for these five items. The internal consistency was .74 in fourth grade, and .74 in fifth grade.

7.7 Peer Social Support

In fourth and fifth grades, children were asked about the social support they received from their peers at school. Six items adapted from the "Kids in My Class at School" questionnaire from the National Institute of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development (SECCYD) (Vandell 2000) were administered. These items were originally from Berndt and Keefe (1995), later adapted by Ladd (e.g., Ladd, Kochendorfer, and Coleman 1996), and later adapted for use in the NICHD SECCYD. There were 10 items in the scale included in the NICHD SECCYD that were rated on a 5-point scale (Never, Hardly ever, Sometimes, Most of the time, Always). Six of these 10 items were administered in the ECLS-K:2011, and children were asked to rate these items on an adapted 5-point scale (Never, Rarely, Sometimes, Often, Very often). During the fourth- and fifth-grade data collections, the child questionnaire included the following six items to measure peer social

support: “Kids in my class make me feel better if I’m having a bad day,” “Kids in my class let me play with them,” “Kids in my class make me feel happy,” “Kids in my class would help me if I hurt myself on the playground,”[5] “Kids in my class tell me that I’m their friend,” and “Kids in my class would help me if other kids were being mean to me.” The peer social support items were ordered in the questionnaire after the behavioral engagement items in fourth grade and after the worry/stress about school items in fifth grade, intermixed with the fear-of-negative-evaluation and loneliness items, and before the peer victimization items.

Cronbach’s alpha was used to examine the internal consistency reliability of these six items. The internal consistency of these items was .84 in fourth grade and .87 in fifth grade.

7.8 Loneliness

Children were asked about their feelings of loneliness at school using three items in the fourth- and fifth-grade data collections. The item wording and response options were adapted from Parker and Asher (1993). The items administered were the following: “I feel lonely at school,”[6] “I feel left out at school,” and “I feel alone at school.” In the original items, children used a 5-point scale to indicate the degree to which each statement was a true description of themselves. In the ECLS-K:2011, children responded using a 5-point frequency scale (Never, Rarely, Sometimes, Often, Very often), which was used for many of the items on the child questionnaire. The loneliness items were intermixed with the fear-of-negative-evaluation and peer social support items, and the same response options were used for all of these items (Never, Rarely, Sometimes, Often, Very often).

Internal consistency reliability (Cronbach’s alpha) was examined for these items. These three items had an internal consistency of .85 in fourth grade and .89 in fifth grade.

7.9 Media Usage

The child questionnaire had three questions on media usage in fourth grade and five questions in fifth grade. One of the questions in fourth grade and three of the questions in fifth grade asked about the frequency of media usage and two questions asked about whether the child’s family had

[5] In fifth grade, the wording of this item was changed to “Kids in my class would help me if I hurt myself at school” to make it more inclusive of all school experiences.
[6] The original item from Parker and Asher (1993) was “I’m lonely at school.”

rules about media usage. The frequency-of-use items were adapted from the Pew September Tracking Survey 2009 that was part of the Pew Internet and American Life Project (Princeton Survey Research Associates International 2009). In fourth and fifth grades, children were asked "How often do you send texts, messages, or emails, using an app, cell phone, computer, iPad or tablet, or other electronic device?" In fifth grade, two additional items were added: "How often do you play an online game against another online player using an app, cell phone, computer, iPad or tablet, or other electronic device?" and "How often do you use a social networking site like Facebook, Twitter, Instagram, or Vine using an app, cell phone, computer, iPad or tablet, or other electronic device?" Children indicated their response using a 5-point scale (Never, Less than once a week, A few times a week, About once a day, Many times a day). In fourth and fifth grades, children were asked the following two questions about rules families have about media use: "Does your family have rules about who you can send texts, messages, or emails to?" and "Does your family have rules about when you can send texts, messages, or emails to?" Children responded "Yes" or "No."

7.10 Pets

At the end of the fourth-grade questionnaire, children were asked a set of questions about family pets adapted from the CENSHARE Pet Attachment Survey (Holcomb, Williams, and Richards 1985). Children were asked whether or not they had a family pet, if they had ever had a pet, how old they were when they got their first pet, how many pets they currently had, and what kind of pet(s) they had. Children who reported at least one pet were also asked a number of questions about that pet (or their favorite pet, if more than one pet) and their relationship with that pet. Children were asked how often they spent time playing with their pet/favorite pet and how often they have their pet near them when they do homework, study, read, or watch TV. In addition, children were asked how often they go to their pet/favorite pet/any of their pets to help them feel better when feeling bad or sad and if they consider their pet/favorite pet/any of their pets to be a member of the family.

7.11 School Belonging

In fifth grade, children were asked five questions about their feelings about school belonging. The questions about school belonging were originally asked in the Grade 8 Student Questionnaire from ECLS-K. Children were asked, "This school year, how often did you…" "feel like you fit in at your school?," "feel close to classmates at your school?," "feel close to teachers in your school?," " enjoy being at your school?," and "feel safe at your school?" Children were asked to indicate

their response on a 4-point scale (Never, Sometimes, Often, Always). These items followed the behavioral engagement items in the fifth-grade questionnaire and came before the items that measured grit.

The internal consistency reliabiliy (Cronbach's alpha) was examined for these items. These items had an internal consistency of .70 in fifth grade.

7.12 Grit

The fifth-grade child questionnaire included six items to assess grit (i.e., sticking with things over the very long term until they are mastered). These items were developed in collaboration with Angela Duckworth for the purpose of this study and are adapted from items in the Character Growth Card (A. Duckworth, personal communication 2015). Children were asked to respond using a 5-point scale (Not at all like me, A little bit like me, Somewhat like me, Quite a bit like me, Very much like me) on the following items: "I finish whatever I begin," "I try hard even after making mistakes," "I continue to work towards my goals, even when they take a long time to complete," "I keep working hard even when I feel like quiting," "I continue working on what I set out to do, even when it takes a long time to complete," and "I try to improve myself, even when it takes a long time to get there." These items were ordered in the questionnaire after the school belonging items.

Internal consistency reliabilty (Cronbach's alpha) was examined for these items in the fifth-grade data collection. The internal consistency was .83 for these six items.

7.13 Worry/Stress About School

The fifth-grade child questionnaire included five questions that asked children about their worry or stress about school. These questions were selected from a larger set of items on internalizing problem behaviors that was developed for and used in grades 3, 5, and 8 of the prior cohort study, ECLS-K. Children were asked "How true is each of these things about you?," and they were asked to respond on a 4-point scale (Not at all true, A little bit true, Mostly true, Very true). The five items asked in the fifth-grade round of data collection were as follows: "I worry about taking tests," "It's hard for me to finsh my school work," "I feel ashamed when I make mistakes at school," "I worry about doing well in school," and "I worry about finishing my work." These items were ordered in the questionnaire after the

behavioral engagement, school belonging, and grit items and immediately before the set of items that included peer social support, fear of negative evaluation, and loneliness.

The internal consistency reliability (Cronbach's alpha) was examined for these items. The internal consistency of these five items was .71.

7.14 Parental Monitoring

In fifth grade, children were asked three items about parental monitoring. These items were taken and adapted from a larger group of nine items administered by Kerr and Stattin (2000). Kerr and Stattin used a 5-point response scale, but it was adapted to a 6-point scale for the ECLS-K:2011 (Never, Rarely, Sometimes, Often, Very often, Always). Children were asked "How often do your parents/guardians…" "Know what you do during your free time," "Know how much homework you have," and "Know what your school grades are." These items were ordered in the questionnaire after the items on media use and family rules about media use and before the life satisfaction items.

Internal consistency reliability (Cronbach's alpha) was examined for these three items. The internal consistency was .56. Users may wish to use these questions as single-item indicators rather than as a scale.

8. PSYCHOMETRIC CHARACTERISTICS OF THE INDIRECT CHILD MEASURES

In the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), teacher and parent reports about children's skills and behaviors are referred to as indirect measures. This chapter describes the indirect measures included in the third-, fourth-, and fifth-grade rounds of data collection. Teacher-reported and parent-reported measures of children's cognitive and socioemotional knowledge and skills and perceptions of relationships with others are described. 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. 2015a) and the *ECLS-K:2011 Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a). For information about the indirect measures used in the first- and second-grade rounds of data collection, see the *User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES 2017-285) (Tourangeau et al. 2017) and the *ECLS-K:2011 First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b).

This chapter includes information about indirect data from the kindergarten through second-grade rounds of data collection in three instances: when those data have been changed since their release on previous files, when new data from those rounds have been added to the kindergarten through fifth-grade (K–5) data file, and when necessary to illustrate how third-, fourth-, and fifth-grade data related to a particular measure or construct differ from data related to the same measure or construct released for the earlier rounds. Information about indirect measures that were used in prior rounds but not in third through fifth grades, for example, parent-reported social skills and Approaches to Learning items administered in fall and spring of kindergarten and spring of first grade, can be found in the following manuals:

- *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. 2015a), hereinafter referred to as the base-year User's Manual;
- *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) (Tourangeau et al. 2015b); and
- *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES 2017-285) (Tourangeau et al. 2017).

8.1 Teacher Measures

This section presents information on the teacher-reported perceptions of the child's skills, behaviors, and relationships. Specifically, it provides information on the teacher's perception of the child's social skills, approaches to learning behaviors, attentional focusing, inhibitory control, working memory, peer context, peer relationships, and school liking behaviors, as well as the student-teacher relationship. This section on teacher measures focuses on child behaviors and relationships reported by teachers in the third-grade, fourth-grade, and fifth-grade data collections.

In kindergarten through third grade, the child's classroom teacher completed a child-level teacher questionnaire that included questions about the child's behavior. A single classroom teacher was asked to report for each child in these earlier grades because it is more typical for a child to have only one teacher or to be taught by one teacher for a majority or significant portion of the day. The ECLS-K:2011 made a major change in its approach to collecting the teacher questionnaire data starting in fourth grade because it becomes increasingly more likely that students would have different teachers for different subjects as students progress through elementary school. In fourth and fifth grades, instead of having a single child-level teacher questionnaire, there were three separate subject-specific child-level teacher questionnaires: one for the child's reading and language arts teacher, one for the child's mathematics teacher, and one for the child's science teacher. (See chapter 2 of the *User's Manual for the ECLS-K:2011 Kindergarten–Fourth Grade Data File and Electronic Codebook, Public Version* [NCES 2018-032] [Tourangeau et al. 2018b] for additional information on the structure of the teacher questionnaires.) The reading, mathematics, and science subject-specific child-level teacher questionnaires each contained classroom-level questions related to the content of the class but also a few child-level questions specifically related to either the child's reading, mathematics, or science experience and one question related to classroom-level social and self-regulatory child behaviors in the specific class. The reading teacher was asked to answer additional child-level questions that were not included in the mathematics and science teacher questionnaires, many of which were asked of the classroom teacher in prior rounds of data collection (kindergarten through third grade), including reports of the teacher's perceptions of the child's behaviors. In fourth and fifth grades, the teacher identified as the child's reading and language arts teacher reported his or her perceptions of the child's behavior, including social skills, approaches to learning, attentional focusing, inhibitory control, school liking, and social interactions and relationships in the classroom.

8.1.1 Teacher-Reported Social Skills

In the fall and spring data collections in kindergarten through second grade, and the spring data collections in third, fourth, and fifth grades, teachers reported how often their ECLS-K:2011 students 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 measuring children's social skills and behaviors are based on items from the *Social Skills Rating System* (Gresham and Elliott 1990)[1] and were included in the self-administered child-level teacher questionnaire in kindergarten, first grade, second grade, and third grade and in the child-level teacher questionnaire for the reading and language arts teacher in fourth and fifth grades. The social skills battery includes some items taken verbatim from the *Social Skills Rating System*, some items that are modifications of original *Social Skills Rating Systems* items, and some items that measure the same kinds of skills and behaviors captured in the *Social Skills Rating System* but use wording developed specifically for the ECLS studies.

Four social skill scales were developed based on teachers' responses to these questionnaire items. 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 (four items), Interpersonal Skills (five items), Externalizing Problem Behaviors (six items),[2] and Internalizing Problem Behaviors (four 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 numbers of items that were required to compute a score were as follows: Self-Control (three out of four items), Interpersonal Skills (four out of five items), Externalizing Problem Behaviors (four out of six items), and Internalizing Problem Behaviors (three out of four items). Higher scores indicate that the child exhibited the behavior represented by the scale more often (e.g., higher Self-Control scores indicate that the child exhibited behaviors indicative of self-control more often; higher Externalizing Problem Behaviors scores indicate that the child exhibited more externalizing behavior problems). Variable names for the teacher scale scores, descriptions, value ranges, weighted means, and standard deviations for these scales are shown in table 8-1.[3]

[1] The *Social Skills Rating System* (SSRS) is an NCS Pearson 1990 copyrighted instrument and has been adapted with permission. These are items developed by Gresham and Elliott (1990).

[2] For children who were in first grade during the first-grade data collections (rounds 3 and 4) and for all children in subsequent rounds of data collection (rounds 5, 6, 7, 8, and 9), the Externalizing Problem Behaviors composite is based on six items. This is different from how the composite was created for the kindergarten rounds (rounds 1 and 2). One additional item was included at the end of the "Social Skills" section of the questionnaire in first, second, third, fourth, and fifth grades. 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.

[3] 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 *ECLS-K:2011 User's Manual for the ECLS-K:2011 Kindergarten–First Grade Data File and Electronic Codebook, Public Version* (NCES 2015-078) (Tourangeau et al. 2015b).

Data for the individual items contributing to each scale were not included in the K–4 data file or any prior data file due to copyright restrictions. The item-level data for all rounds of data collection appear for the first time on the K–5 data file because permission to include them was granted from the publisher.

Table 8-1. Teacher-reported social skills scales variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1TCHCON	X1 Teacher Report Self-Control	13,550	1–4	3.07	0.629
X1TCHPER	X1 Teacher Report Interpersonal Skills	13,708	1–4	2.98	0.639
X1TCHEXT	X1 Teacher Report Externalizing Problem Behaviors	14,385	1–4	1.61	0.631
X1TCHINT	X1 Teacher Report Internalizing Problem Behaviors	14,239	1–4	1.47	0.494
X2TCHCON	X2 Teacher Report Self-Control	15,796	1–4	3.17	0.637
X2TCHPER	X2 Teacher Report Interpersonal Skills	15,799	1–4	3.13	0.650
X2TCHEXT	X2 Teacher Report Externalizing Problem Behaviors	15,903	1–4	1.64	0.639
X2TCHINT	X2 Teacher Report Internalizing Problem Behaviors	15,865	1–4	1.51	0.498
X3TCHCON	X3 Teacher Report Self-Control	4,658	1–4	3.21	0.591
X3TCHPER	X3 Teacher Report Interpersonal Skills	4,724	1–4	3.14	0.613
X3TCHEXT	X3 Teacher Report Externalizing Problem Behaviors	4,964	1–4	1.67	0.590
X3TCHINT	X3 Teacher Report Internalizing Problem Behaviors	4,848	1–4	1.48	0.483
X4TCHCON	X4 Teacher Report Self-Control	13,202	1–4	3.21	0.621
X4TCHPER	X4 Teacher Report Interpersonal Skills	13,288	1–4	3.14	0.657
X4TCHEXT	X4 Teacher Report Externalizing Problem Behaviors	13,398	1–4	1.73	0.619
X4TCHINT	X4 Teacher Report Internalizing Problem Behaviors	13,306	1–4	1.55	0.508
X4KTCHCON	X4K Teacher Report Self-Control	418	1–4	3.09	0.616
X4KTCHPER	X4K Teacher Report Interpersonal Skills	418	1–4	3.04	0.671
X4KTCHEXT	X4K Teacher Report Externalizing Problem Behaviors	419	1–4	1.78	0.614
X4KTCHINT	X4K Teacher Report Internalizing Problem Behaviors	418	1–4	1.62	0.498

See notes at end of table.

Table 8-1. Teacher-reported social skills scales variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X5TCHCON	X5 Teacher Report Self-Control	4,174	1–4	3.23	0.614
X5TCHPER	X5 Teacher Report Interpersonal Skills	4,178	1–4	3.13	0.621
X5TCHEXT	X5 Teacher Report Externalizing Problem Behaviors	4,426	1–4	1.65	0.610
X5TCHINT	X5 Teacher Report Internalizing Problem Behaviors	4,342	1–4	1.50	0.522
X6TCHCON	X6 Teacher Report Self-Control	12,472	1–4	3.22	0.629
X6TCHPER	X6 Teacher Report Interpersonal Skills	12,518	1–4	3.12	0.664
X6TCHEXT	X6 Teacher Report Externalizing Problem Behaviors	12,657	1–4	1.72	0.625
X6TCHINT	X6 Teacher Report Internalizing Problem Behaviors	12,577	1–4	1.59	0.528
X7TCHCON	X7 Teacher Report Self-Control	11,736	1–4	3.27	0.619
X7TCHPER	X7 Teacher Report Interpersonal Skills	11,768	1–4	3.14	0.657
X7TCHEXT	X7 Teacher Report Externalizing Problem Behaviors	11,898	1–4	1.69	0.615
X7TCHINT	X7 Teacher Report Internalizing Problem Behaviors	11,830	1–4	1.61	0.535
X8TCHCON	X8 Teacher Report Self-Control	10,848	1–4	3.28	0.604
X8TCHPER	X8 Teacher Report Interpersonal Skills	10,867	1–4	3.12	0.648
X8TCHEXT	X8 Teacher Report Externalizing Problem Behaviors	11,000	1–4	1.65	0.594
X8TCHINT	X8 Teacher Report Internalizing Problem Behaviors	10,923	1–4	1.58	0.534
X9TCHCON	X9 Teacher Report Self-Control	10,235	1–4	3.29	0.609
X9TCHPER	X9 Teacher Report Interpersonal Skills	10,224	1–4	3.13	0.650
X9TCHEXT	X9 Teacher Report Externalizing Problem Behaviors	10,359	1–4	1.63	0.590
X9TCHINT	X9 Teacher Report Internalizing Problem Behaviors	10,294	1–4	1.57	0.518

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P3T0, and spring first-grade estimates (X4) are weighted by W4CS4P_2T0. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_2T0. Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. Items contributing to the teacher-reported social skill scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. Variables that begin with "X4K" are for data collected in the spring first-grade data collection for children who were retained in kindergarten. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.
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, spring 2013, spring 2014, spring 2015, and spring 2016.

Table 8-2 presents the internal consistency reliability (Cronbach's alpha) estimates of the Self-Control, Interpersonal Skills, Externalizing Problem Behaviors, and Internalizing Problem Behaviors scales derived from information reported by the teacher.

Table 8-2. Teacher-reported social skills scales reliability estimates for fall and spring kindergarten, fall and spring first grade, and fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Variable name	Description	Number of items	Reliability coefficient
X1TCHCON	X1 Teacher Report Self-Control	4	.81
X1TCHPER	X1 Teacher Report Interpersonal Skills	5	.86
X1TCHEXT	X1 Teacher Report Externalizing Problem Behaviors	5	.88
X1TCHINT	X1 Teacher Report Internalizing Problem Behaviors	4	.79
X2TCHCON	X2 Teacher Report Self-Control	4	.82
X2TCHPER	X2 Teacher Report Interpersonal Skills	5	.87
X2TCHEXT	X2 Teacher Report Externalizing Problem Behaviors	5	.89
X2TCHINT	X2 Teacher Report Internalizing Problem Behaviors	4	.78
X3TCHCON	X3 Teacher Report Self-Control	4	.79
X3TCHPER	X3 Teacher Report Interpersonal Skills	5	.85
X3TCHEXT	X3 Teacher Report Externalizing Problem Behaviors	5	.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	5	.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	5	.87
X4KTCHINT	X4K Teacher Report Internalizing Problem Behaviors	4	.73

See notes at end of table.

Table 8-2. Teacher-reported social skills scales reliability estimates for fall and spring kindergarten, fall and spring first grade, and fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Variable name	Description	Number of items	Reliability coefficient
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
X7TCHCON	X7 Teacher Report Self-Control	4	.80
X7TCHPER	X7 Teacher Report Interpersonal Skills	5	.86
X7TCHEXT	X7 Teacher Report Externalizing Problem Behaviors	6	.87
X7TCHINT	X7 Teacher Report Internalizing Problem Behaviors	4	.78
X8TCHCON	X8 Teacher Report Self-Control	4	.80
X8TCHPER	X8 Teacher Report Interpersonal Skills	5	.86
X8TCHEXT	X8 Teacher Report Externalizing Problem Behaviors	6	.87
X8TCHINT	X8 Teacher Report Internalizing Problem Behaviors	4	.79
X9TCHCON	X9 Teacher Report Self-Control	4	.80
X9TCHPER	X9 Teacher Report Interpersonal Skills	5	.86
X9TCHEXT	X9 Teacher Report Externalizing Problem Behaviors	6	.88
X9TCHINT	X9 Teacher Report Internalizing Problem Behaviors	4	.79

NOTE: Items contributing to the teacher-reported social skill scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. Variables that begin with "X4K" are for data collected in the spring first-grade data collection for children who were retained in kindergarten.
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, spring 2013, spring 2014, spring 2015, and spring 2016.

Within-round correlations among the four social skills scales (Self-Control, Interpersonal Skills, Externalizing Problem Behaviors, and Internalizing Problem Behaviors) for third, fourth and fifth grades (rounds 7-9) are presented in table 8-3. See chapter 7 of the *ECLS-K:2011 First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian et al. 2018b) for the within-round correlations for these variables in kindergarten through second grades (rounds 1-6). 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 1-9. 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 lower ratings of self-control and interpersonal skills. The correlation between Externalizing Problem Behaviors and Self-Control ranges from -.74 to -.68 across rounds 1-9; the correlation between Externalizing Problem Behaviors and Interpersonal Skills ranges from -.63 to -.57 across rounds 1-9. Internalizing Problem Behaviors is negatively associated with Self-Control and Interpersonal Skills, ranging from -.36 to -.26 and from -.38 to -.31, respectively, across rounds 1-9. Externalizing Problem Behaviors is positively related to Internalizing Problem Behaviors, with correlations ranging from .26 to .33 across rounds 1-9.

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 8-3. Within-round correlations of teacher-reported social skills for spring third grade, spring fourth grade, and spring fifth grade: Spring 2014, spring 2015, and spring 2016

Round	Self-control (X*TCHCON)	Interpersonal skills (X*TCHPER)	Externalizing behavior (X*TCHEXT)	Internalizing behavior (X*TCHINT)
Round 7: Spring third grade				
Self-control	1.00	†	†	†
Interpersonal skills	.80	1.00	†	†
Externalizing behavior	-.73	-.62	1.00	†
Internalizing behavior	-.34	-.37	. 32	1.00
Round 8: Spring fourth grade				
Self-control	1.00	†	†	†
Interpersonal skills	.80	1.00	†	†
Externalizing behavior	-.73	-.62	1.00	†
Internalizing behavior	-.34	-.36	.31	1.00
Round 9: Spring fifth grade				
Self-control	1.00	†	†	†
Interpersonal skills	.80	1.00	†	†
Externalizing behavior	-.73	-.62	1.00	†
Internalizing behavior	-.34	-.37	.31	1.00

† Not applicable.

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 respondent in third grade (round 7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The "*" in the variables names is to be substituted for the data collection round number (7 to 9). All estimates unweighted. All correlations $p < .0001$.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014, spring 2015, and spring 2016.

Mean scores for the teacher-reported social skills subscales are presented by data collection round and child characteristics in tables 8-4 through 8-7.

Table 8-4. Mean teacher-reported Self-Control score, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

	Spring third grade (X7TCHCON)			Spring fourth grade (X8TCHCON)			Spring fifth grade (X9TCHCON)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**11,736**	**3.27**	**0.619**	**10,848**	**3.28**	**0.604**	**10,235**	**3.29**	**0.609**
Sex									
Male	6,006	3.15	0.648	5,560	3.16	0.632	5,229	3.17	0.638
Female	5,715	3.39	0.562	5,274	3.40	0.545	4,996	3.42	0.549
Race/ethnicity									
White, non-Hispanic	5,738	3.30	0.605	5,370	3.33	0.585	5,081	3.35	0.593
Black, non-Hispanic	1,283	3.01	0.684	1,164	3.00	0.656	968	2.99	0.655
Hispanic	3,160	3.31	0.589	2,901	3.29	0.584	2,849	3.30	0.587
Asian, non-Hispanic	889	3.40	0.547	814	3.38	0.536	781	3.41	0.557
Hawaiian, Other Pacific Islander, non-Hispanic	66	3.39	0.631	52	3.37	0.664	49	3.39	0.570
American Indian/Alaska Native, non-Hispanic	100	3.31	0.527	96	3.35	0.469	94	3.42	0.535
Two or more races, non-Hispanic	489	3.24	0.637	442	3.29	0.620	404	3.35	0.581
School type									
Public school	10,624	3.26	0.621	9.842	3.27	0.607	9,290	3.29	0.613
Private school	1,112	3.33	0.593	1,006	3.34	0.562	945	3.29	0.570

NOTE: Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 4. Detail may not sum to totals due to rounding and/or missing data. *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 2014, spring 2015, and spring 2016.

Table 8-5. Mean teacher-reported Interpersonal Skills score, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

Characteristic	Spring third grade (X7TCHPER)			Spring fourth grade (X8TCHPER)			Spring fifth grade (X9TCHPER)		
	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**11,768**	**3.14**	**0.657**	**10,867**	**3.12**	**0.648**	**10,224**	**3.13**	**0.650**
Sex									
Male	6,009	2.99	0.662	5,537	2.98	0.660	5,203	2.98	0.657
Female	5,744	3.29	0.616	5,316	3.27	0.602	5,011	3.28	0.605
Race/ethnicity									
White, non-Hispanic	5,759	3.18	0.649	5,402	3.17	0.641	5,085	3.18	0.646
Black, non-Hispanic	1,285	2.91	0.680	1,149	2.89	0.660	965	2.87	0.665
Hispanic	3,159	3.16	0.642	2,904	3.11	0.633	2,842	3.13	0.628
Asian, non-Hispanic	900	3.20	0.618	812	3.20	0.618	775	3.14	0.669
Hawaiian, Other Pacific Islander, non-Hispanic	66	3.15	0.751	53	3.27	0.660	49	3.18	0.571
American Indian/Alaska Native, non-Hispanic	98	3.22	0.570	97	3.19	0.562	95	3.30	0.544
Two or more races, non-Hispanic	490	3.12	0.675	441	3.11	0.677	404	3.19	0.619
School type									
Public school	10,666	3.12	0.658	9,861	3.11	0.650	9,277	3.12	0.653
Private school	1,102	3.27	0.623	1,006	3.23	0.614	947	3.18	0.620

NOTE: Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 4. Detail may not sum to totals due to rounding and/or missing data. *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 2014, spring 2015, and spring 2016.

Table 8-6. Mean teacher-reported Externalizing Problem Behaviors score, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

Characteristic	Spring third grade (X7TCHEXT)			Spring fourth grade (X8TCHEXT)			Spring fifth grade (X9TCHEXT)		
	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**11,898**	**1.69**	**0.615**	**11,000**	**1.65**	**0.594**	**10,359**	**1.63**	**0.590**
Sex									
Male	6,090	1.83	0.658	5,627	1.79	0.635	5,291	1.77	0.624
Female	5,793	1.54	0.525	5,359	1.50	0.504	5,058	1.48	0.511
Race/ethnicity									
White, non-Hispanic	5,797	1.66	0.591	5,440	1.62	0.574	5,133	1.61	0.569
Black, non-Hispanic	1,302	1.93	0.702	1,171	1.91	0.673	977	1.91	0.696
Hispanic	3,210	1.63	0.597	2,957	1.60	0.568	2,899	1.57	0.559
Asian, non-Hispanic	918	1.51	0.482	827	1.47	0.477	789	1.49	0.504
Hawaiian, Other Pacific Islander, non-Hispanic	66	1.66	0.573	54	1.64	0.553	49	1.55	0.511
American Indian/Alaska Native, non-Hispanic	100	1.54	0.493	98	1.48	0.351	95	1.55	0.460
Two or more races, non-Hispanic	494	1.76	0.657	444	1.70	0.647	408	1.62	0.534
School type									
Public school	10,784	1.69	0.619	9,990	1.65	0.595	9,404	1.63	0.591
Private school	1,114	1.67	0.566	1,010	1.63	0.572	955	1.66	0.581

NOTE: Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 4. Detail may not sum to totals due to rounding and/or missing data. *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 2014, spring 2015, and spring 2016.

Table 8-7. Mean teacher-reported Internalizing Problem Behaviors score, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

Characteristic	Spring third grade (X7TCHINT)			Spring fourth grade (X8TCHINT)			Spring fifth grade (X9TCHINT)		
	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**11,830**	**1.61**	**0.535**	**10,923**	**1.58**	**0.534**	**10,294**	**1.57**	**0.518**
Sex									
Male	6,059	1.62	0.545	5,582	1.59	0.551	5,255	1.57	0.526
Female	5,756	1.59	0.525	5,328	1.57	0.515	5,029	1.56	0.510
Race/ethnicity									
White, non-Hispanic	5,796	1.60	0.543	5,422	1.58	0.535	5,123	1.57	0.520
Black, non-Hispanic	1,283	1.67	0.558	1,154	1.62	0.559	964	1.57	0.504
Hispanic	3,175	1.58	0.497	2,927	1.56	0.511	2,871	1.55	0.519
Asian, non-Hispanic	909	1.50	0.481	819	1.51	0.497	777	1.52	0.507
Hawaiian, Other Pacific Islander, non-Hispanic	66	1.48	0.451	53	1.40	0.321	49	1.43	0.484
American Indian/Alaska Native, non-Hispanic	100	1.72	0.610	98	1.59	0.548	93	1.54	0.463
Two or more races, non-Hispanic	491	1.68	0.588	441	1.66	0.600	408	1.62	0.558
School type									
Public school	10,715	1.61	0.540	9,916	1.58	0.535	9,340	1.57	0.519
Private school	1,115	1.55	0.472	1,007	1.56	0.514	954	1.56	0.504

NOTE: Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to the teacher-reported social skills scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 4. Detail may not sum to totals due to rounding and/or missing data. *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 2014, spring 2015, and spring 2016.

8.1.2 Teacher-Reported Approaches to Learning Items and Scale

The child-level teacher questionnaire fielded in every round of data collection from the fall of kindergarten to the spring of third grade and the child-level reading and language arts teacher subject-specific child-level teacher questionnaire in fourth and fifth grades included seven items, referred to as "Approaches to Learning" items, that asked the teachers to report how often their ECLS-K:2011 students 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).[4] These items were presented in the same item set as the social skills items adapted from the *Social Skills Rating System* (described above in section 8.1.1), and teachers used the same frequency scale to report how often each child demonstrated the behaviors described. The Approaches to Learning scale score is the mean rating on the seven items included in the scale. A score was computed when the respondent provided a rating on at least four of the seven items that composed the scale. Higher scale scores indicate that the child exhibited positive learning behaviors more often. The 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 and spring kindergarten child-level teacher questionnaire begin with "T1" and "T2," respectively. 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." Variable names for the spring of third grade begin with "T7," and those for spring of fourth grade begin with "G8." Variable names for the spring of fifth grade begin with "G9." The variable names, descriptions, value ranges, weighted means, and standard deviations for the teacher-reported Approaches to Learning scale scores are shown in table 8-8. The Approaches to Learning scale has a reliability estimate of .91 for each round of data collection from kindergarten through fourth grade and .92 for fifth grade, as measured by Cronbach's alpha.

[4] The Approaches to Learning teacher items were developed specifically for the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K); they were not taken from an existing source. These items 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 used in the kindergarten rounds of the ECLS-K. The seventh item (i.e., follows classroom rules) was added in the first-grade round of the ECLS-K.

Table 8-8. Teacher-reported Approaches to Learning scale variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1TCHAPP	X1 Teacher Report Approaches to Learning	14,770	1–4	2.93	0.680
X2TCHAPP	X2 Teacher Report Approaches to Learning	15,978	1–4	3.09	0.689
X3TCHAPP	X3 Teacher Report Approaches to Learning	5,022	1–4	3.04	0.677
X4TCHAPP	X4 Teacher Report Approaches to Learning	13,449	1–4	3.07	0.700
X4KTCHAPP	X4K Teacher Report Approaches to Learning	417	1–4	2.94	0.704
X5TCHAPP	X5 Teacher Report Approaches to Learning	4,507	1–4	3.05	0.688
X6TCHAPP	X6 Teacher Report Approaches to Learning	12,689	1–4	3.07	0.707
X7TCHAPP	X7 Teacher Report Approaches to Learning	11,913	1–4	3.08	0.711
X8TCHAPP	X8 Teacher Report Approaches to Learning	11,028	1–4	3.09	0.696
X9TCHAPP	X9 Teacher Report Approaches to Learning	10,403	1–4	3.11	0.696

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimates (X2) are weighted by W1C0. Fall first-grade estimates (X3) are weighted by W3CF3P3T0, and spring first-grade estimates (X4) are weighted by W4CS4P_2T0. Fall second-grade estimates (X5) are weighted by W6CF6P_2A0, and spring second-grade estimates (X6) are weighted by W6CS6P_2T0. Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. Variables that begin with "X4K" are for data collected in the spring first-grade data collection for children who were retained in kindergarten. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.
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, spring 2013, spring 2014, spring 2015, and spring 2016.

Within-round correlations between teacher-reported Approaches to Learning and teacher-reported social skills were examined for all rounds of data collection (kindergarten through fifth grade) and are presented in table 8-9. All within-round correlations were significant across all rounds. Approaches to Learning was positively correlated with Interpersonal Skills and Self-Control within rounds, with correlations ranging from .71 to .74 for Interpersonal Skills and from .68 to .72 for Self-Control. There were significant negative within-round correlations between Approaches to Learning and Externalizing Problem Behaviors and between Approaches to Learning and Internalizing Problem Behaviors. Across all rounds, Externalizing Problem Behaviors was negatively associated with Approaches to Learning (range -.62 to -.58), and Internalizing Problem Behaviors was also negatively associated with Approaches to Learning (range -.42 to -.30).

Table 8-9. Within-round correlations of the teacher-reported Approaches to Learning scale and the teacher-reported social skills scales for fall and spring kindergarten, fall and spring first grade, fall and spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School years 2010–11, 2011–12, and 2012–13; spring 2014; spring 2015; and spring 2016

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
Round 7: X7TCHAPP	.70	.72	-.62	-.41
Round 8: X8TCHAPP	.71	.73	-.62	-.40
Round 9: X9TCHAPP	.72	.73	-.62	-.40

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 respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The "*" in the variables names is to be substituted for the data collection round number (1 to 9). All estimates unweighted. All correlations $p < .0001$.
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, spring 2013, spring 2014, spring 2015, and spring 2016.

Mean scores for the teacher-reported Approaches to Learning scale are presented by data collection round and child characteristics in table 8-10.

Table 8-10. Mean teacher-reported Approaches to Learning scale scores, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

Characteristic	Spring third grade (X7TCHAPP)			Spring fourth grade (X8TCHAPP)			Spring fifth grade (X9TCHAPP)		
	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**11,913**	**3.08**	**0.711**	**11,028**	**3.09**	**0.696**	**10,403**	**3.11**	**0.696**
Sex									
Male	6,095	2.91	0.717	5,641	2.90	0.706	5,311	2.92	0.707
Female	5,803	3.26	0.658	5,373	3.29	0.627	5,082	3.32	0.624
Race/ethnicity									
White, non-Hispanic	5,805	3.13	0.693	5,451	3.15	0.677	5,149	3.18	0.683
Black, non-Hispanic	1,303	2.80	0.729	1,172	2.84	0.718	982	2.83	0.698
Hispanic	3,216	3.08	0.706	2,971	3.05	0.704	2,916	3.08	0.695
Asian, non-Hispanic	918	3.33	0.648	828	3.33	0.591	793	3.33	0.644
Hawaiian, Other Pacific Islander, non-Hispanic	66	3.13	0.729	54	3.20	0.682	49	3.34	0.732
American Indian/Alaska Native, non-Hispanic	100	3.12	0.725	98	3.18	0.617	95	3.11	0.664
Two or more races, non-Hispanic	494	3.05	0.722	445	3.12	0.718	410	3.17	0.662
School type									
Public school	10,798	3.07	0.716	10,016	3.08	0.701	9,448	3.11	0.699
Private school	1,115	3.21	0.634	1,012	3.18	0.639	955	3.13	0.667

NOTE: Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to the teacher-reported social skill scales were adapted with permission from the *Social Skills Rating System* (SSRS) (©1990 NCS Pearson). The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 4. Detail may not sum to totals due to rounding and/or missing data. *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 2014, spring 2015, and spring 2016.

8.1.3 Teacher-Reported Attentional Focusing and Inhibitory Control: Children's Behavior Questionnaire (CBQ) and Temperament in Middle Childhood Questionnaire (TMCQ)

The fall kindergarten, spring kindergarten, and 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* (Putnam and Rothbart 2006)[5] asking teachers to indicate how often their ECLS-K:2011 children exhibited certain social skills and behaviors related to inhibitory control and attentional focusing, two indicators related to executive

[5] The *Children's Behavior Questionnaire* is a copyrighted instrument: 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.

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* (*TMCQ*) (Simonds and Rothbart 2004)[6] in the spring of second grade. The *TMCQ* was designed as an upward-age extension of the *CBQ* and is appropriate for children 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, although 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 ("Almost always untrue," "Usually untrue," "Sometimes true/Sometimes untrue," "Usually true," "Almost always true"). Like the *CBQ*, there was a "not applicable" option that the teacher could select if the statement or situation did not apply to the child.

Item-level data for the items that make up the Attentional Focusing and Inhibitory Control scales are provided on the K–5 data file. Variable names for the item-level data from the fall and spring kindergarten child-level teacher questionnaire 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 during spring 2012 begin with "T4K." Variable names for the spring second grade begin with "T6," and those for spring third grade begin with "T7." Variable names from the reading subject-specific child-level questionnaire begin with "G8" for the spring of fourth grade and "G9" for the spring of fifth grade.

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

The data file includes two scale scores for each round of data collection in which each measure was included: (1) Attentional Focus and (2) Inhibitory Control. In kindergarten and first grade these scores are derived from the *CBQ*, and in second, third, fourth, and fifth grades these scores are derived primarily from the *TMCQ*, as explained further below. The scale scores were developed using guidelines from the developers of both the *CBQ* and *TMCQ*.

In kindergarten and first grade, the ECLS-K:2011 fielded all 6 items from the Attentional Focusing subscale and all 6 items from the Inhibitory Control subscale of the *CBQ Short Form*. As such, the kindergarten and 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 slightly for use in the ECLS-K:2011 to make the items more appropriate for a school setting.

In second, third, fourth, and fifth grade, the ECLS-K:2011 fielded six of the seven items from the original *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-, third-, fourth-, and fifth-grade Attentional Focusing scale scores reflect the six items fielded by the ECLS-K:2011, not the full set of items in the original *TMCQ* scale. The second-, third-, fourth-, and fifth-grade Inhibitory Control scale scores reflect 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 *TMCQ* Attentional Focusing subscale was modified slightly to make it more appropriate for a school setting and, similarly, one item on the *TMCQ* Inhibitory Control subscale was modified.

For the kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade Attentional 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. Higher scale scores on the Attentional Focus scale indicate that the child exhibited more behaviors that demonstrate the ability to focus attention on cues in the environment that are relevant to the task. 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 necessary for a particular situation. The variable names, descriptions, value ranges, weighted means, and standard deviations for these scales are shown in tables 8-11 and 8-12.

Table 8-11. Children's Behavior Questionnaire variable names, descriptions, value ranges, weighted means, and standard deviations for fall and spring kindergarten and spring first grade: School year 2010–11 and spring 2012

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X1ATTNFS	X1 Teacher Report Attentional Focus	14,562	1–7	4.68	1.323
X1INBCNT	X1 Teacher Report Inhibitory Control	14,556	1–7	4.88	1.291
X2ATTNFS	X2 Teacher Report Attentional Focus	15,937	1–7	4.90	1.329
X2INBCNT	X2 Teacher Report Inhibitory Control	15,925	1–7	5.06	1.292
X4ATTNFS	X4 Teacher Report Attentional Focus	13,390	1–7	4.84	1.292
X4INBCNT	X4 Teacher Report Inhibitory Control	13,399	1–7	5.04	1.287
X4KATTNFS	X4K Teacher Report Attentional Focus	417	1–7	4.61	1.323
X4KINBCNT	X4K Teacher Report Inhibitory Control	417	1–7	4.88	1.267

NOTE: Fall kindergarten estimates (X1) and spring kindergarten estimates (X2) are weighted by W1C0. Spring first-grade estimates (X4) are weighted by W4CS4P_2T0. Items contributing to these scales come from the *Children's Behavior Questionnaire* (Putnam and Rothbart 2006). The respondent in kindergarten and first grade (rounds 1-4) was the child's classroom teacher. Variables that begin with "X4K" are for data collected in the spring first-grade data collection for children who were retained in kindergarten. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.
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, and spring 2012.

Table 8-12. Temperament in Middle Childhood Questionnaire variable names, descriptions, value ranges, weighted means, and standard deviations for spring second grade, spring third grade, spring fourth grade, and spring fifth grade: Spring 2013, spring 2014, spring 2015, and spring 2016

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X6ATTMCQ	X6 TMCQ Teacher Report Attentional Focus	12,661	1–5	3.47	1.122
X6INTMCQ	X6 TMCQ Teacher Report Inhibitory Control	12,659	1–5	3.67	0.845
X7ATTMCQ	X7 TMCQ Teacher Report Attentional Focus	11,879	1–5	3.48	1.119
X7INTMCQ	X7 TMCQ Teacher Report Inhibitory Control	11,882	1–5	3.69	0.825
X8ATTMCQ	X8 TMCQ Teacher Report Attentional Focus	11,008	1–5	3.54	1.112
X8INTMCQ	X8 TMCQ Teacher Report Inhibitory Control	11,002	1–5	3.73	0.812
X9ATTMCQ	X9 TMCQ Teacher Report Attentional Focus	10,367	1–5	3.61	1.083
X9INTMCQ	X9 TMCQ Teacher Report Inhibitory Control	10,355	1–5	3.80	0.802

NOTE: Spring second-grade estimates (X6) are weighted by W6CS6P_2T0. Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. 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 respondent in second grade and third grade (rounds 6 and 7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.
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, spring 2014, spring 2015, and spring 2016.

Table 8-13 presents the internal consistency reliability coefficients (Cronbach's alpha) for the teacher-reported Attentional Focus and Inhibitory Control scales for kindergarten through fifth grade. For the spring of third grade, the Attentional Focus scale (X7ATTMCQ[7]) has an internal consistency reliability coefficient of .96, and the Inhibitory Control scale (X7INTMCQ[8]) has an internal consistency reliability coefficient of .85. In the spring of fourth grade, the internal consistency reliability coefficient is .96 for the Attentional Focus scale (X8ATTMCQ) and .85 for the Inhibitory Control scale (X8INTMCQ). In the spring of fifth grade, the internal consistency reliability coefficient is .96 for the Attentional Focus scale (X9ATTMCQ) and .85 for the Inhibitory Control scale (X9INTMCQ).

The study received copyright permission to include item-level data from both the *CBQ* and the *TMCQ* in the ECLS-K:2011 data files. Therefore, these data have been included in the kindergarten through fifth-grade data file with the other child-level teacher questionnaire data. Variable names for the item-level data from the fall of kindergarten, the spring of kindergarten, the spring of first grade, the spring of second grade, and the spring of third grade begin with "T1," "T2," "T4," "T6," and "T7," respectively. Variable names from the item-level data begin with "G8" for the spring of fourth grade and "G9" for the spring of fifth grade. Variable names that begin with "T4K" are for item-level data from the spring of first grade for students retained in kindergarten in spring 2012.

[7] The variable name for the Attentional Focus composite was changed from X*ATTNFS to X*ATTMCQ starting in second grade. Although the construct is believed to be the same, the items used to derive the composite were from the *CBQ* for kindergarten and first grade but were from the *TMCQ* starting at second grade. Thus, the name of the composite variable was changed.

[8] The variable name for the Inhibitory Control composite was changed from X*INBCNT to X*INTMCQ starting in second grade. Although the construct is believed to be the same, the items used to derive the composite were from the *CBQ* for kindergarten and first grade but were from the *TMCQ* starting at second grade. Thus, the name of the composite variable was changed.

Table 8-13. Reliability estimates for the teacher-reported Attentional Focus and Inhibitory Control scales for fall and spring kindergarten, spring first grade, spring second grade, spring third grade, spring fourth grade, and spring fifth grade: School year 2010–11, spring 2012, spring 2013, spring 2014, spring 2015, and spring 2016

Variable name	Description	Number of items	Reliability coefficient
X1ATTNFS	X1 Teacher Report Attentional Focus	6	.87
X1INBCNT	X1 Teacher Report Inhibitory Control	6	.87
X2ATTNFS	X2 Teacher Report Attentional Focus	6	.87
X2INBCNT	X2 Teacher Report Inhibitory Control	6	.87
X4ATTNFS	X4 Teacher Report Attentional Focus	6	.83
X4INBCNT	X4 Teacher Report Inhibitory Control	6	.86
X4KATTNFS	X4K Teacher Report Attentional Focus	6	.86
X4KINBCNT	X4K Teacher Report Inhibitory Control	6	.86
X6ATTMCQ	X6 TMCQ Teacher Report Attentional Focus	6	.96
X6INTMCQ	X6 TMCQ Teacher Report Inhibitory Control	7	.87
X7ATTMCQ	X7 TMCQ Teacher Report Attentional Focus	6	.96
X7INTMCQ	X7 TMCQ Teacher Report Inhibitory Control	7	.85
X8ATTMCQ	X8 TMCQ Teacher Report Attentional Focus	6	.96
X8INTMCQ	X8 TMCQ Teacher Report Inhibitory Control	7	.85
X9ATTMCQ	X9 TMCQ Teacher Report Attentional Focus	6	.96
X9INTMCQ	X9 TMCQ Teacher Report Inhibitory Control	7	.85

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 respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. Variables that begin with "X4K" are for data collected in the spring first-grade data collection for children who were retained in kindergarten.
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, spring 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

Within-round and across-round correlations for the Attentional Focusing and Inhibitory Control scores produced from the *TMCQ* in third, fourth, and fifth grades are presented in table 8-14.

Mean scores for the teacher-reported Attentional Focus and Inhibitory Control scales are presented by child characteristics in tables 8-15 and 8-16.

Table 8-14. Intercorrelations among Attentional Focusing and Inhibitory Control scales for spring third grade, spring fourth grade, and spring fifth grade: Spring 2014, spring 2015, and spring 2016

	Spring third grade		Spring fourth grade		Spring fifth grade	
Measures	Attentional Focus (X7ATTMCQ)	Inhibitory Control (X7INTMCQ)	Attentional Focus (X8ATTMCQ)	Inhibitory Control (X8INTMCQ)	Attentional Focus (X9ATTMCQ)	Inhibitory Control (X9INTMCQ)
X7ATTMCQ	1.00	†	†	†	†	†
X7INTMCQ	.76	1.00	†	†	†	†
X8ATTMCQ	.61	.54	1.00	†	†	†
X8INTMCQ	.52	.55	.76	1.00	†	†
X9ATTMCQ	.58	.52	.61	.54	1.00	†
X9INTMCQ	.50	.52	.53	.54	.76	1.00

† Not applicable.
NOTE: Items contributing to these scales come from the *Temperament in Middle Childhood Questionnaire* (Simonds and Rothbart 2004). The respondent in third grade (round 7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. All estimates unweighted. All correlations $p < .0001$.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014, spring 2015, and spring 2016.

Table 8-15. Mean Temperament in Middle Childhood Questionnaire Attentional Focus scores, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

	Spring third grade (X7ATTMCQ)			Spring fourth grade (X8ATTMCQ)			Spring fifth grade (X9ATTMCQ)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**11,879**	**3.48**	**1.119**	**11,008**	**3.54**	**1.112**	**10,367**	**3.61**	**1.083**
Sex									
Male	6,077	3.21	1.116	5,627	3.24	1.117	5,297	3.32	1.085
Female	5,787	3.77	1.047	5,367	3.85	1.015	5,060	3.92	0.994
Race/ethnicity									
White, non-Hispanic	5,789	3.53	1.103	5.446	3.58	1.107	5,129	3.68	1.077
Black, non-Hispanic	1,296	3.15	1.121	1,170	3.22	1.076	980	3.25	1.085
Hispanic	3,213	3.48	1.123	2,960	3.52	1.119	2,906	3.59	1.070
Asian, non-Hispanic	912	3.91	0.968	825	3.92	0.981	791	3.97	1.004
Hawaiian, Other Pacific Islander, non-Hispanic	65	3.66	1.129	54	3.86	1.002	49	3.64	0.924
American Indian/Alaska Native, non-Hispanic	100	3.58	1.111	98	3.76	1.045	95	3.78	1.016
Two or more races, non-Hispanic	493	3.46	1.180	446	3.55	1.168	408	3.67	1.052
School type									
Public school	10,768	3.47	1.125	9,995	3.52	1.117	9,419	3.61	1.089
Private school	1,111	3.60	1.042	1,013	3.71	1.032	948	3.67	1.014

NOTE: Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to these scales come from the *Temperament in Middle Childhood Questionnaire* (Simonds and Rothbart 2004). The respondent in third grade (round 7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 5. Detail may not sum to totals due to rounding and/or missing data. *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 2014, spring 2015, and spring 2016.

Table 8-16. Mean Temperament in Middle Childhood Questionnaire Inhibitory Control scores, by data collection round and child characteristics: Spring 2014, spring 2015, and spring 2016

Characteristic	Spring third grade (X7INTMCQ)			Spring fourth grade (X8INTMCQ)			Spring fifth grade (X9INTMCQ)		
	Number	Mean	*SD*	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**11,882**	**3.69**	**0.825**	**11,002**	**3.73**	**0.812**	**10,355**	**3.80**	**0.802**
Sex									
Male	6,078	3.45	0.833	5,623	3.49	0.816	5,291	3.56	0.797
Female	5,789	3.95	0.736	5,365	3.99	0.722	5,054	4.05	0.727
Race/ethnicity									
White, non-Hispanic	5,787	3.73	0.832	5,441	3.78	0.816	5,126	3.85	0.818
Black, non-Hispanic	1,299	3.41	0.813	1,172	3.47	0.778	979	3.53	0.768
Hispanic	3,212	3.72	0.795	2,960	3.72	0.804	2,901	3.79	0.771
Asian, non-Hispanic	914	3.94	0.743	823	4.00	0.742	789	4.02	0.744
Hawaiian, Other Pacific Islander, non-Hispanic	66	3.77	0.785	54	3.86	0.857	49	3.78	0.710
American Indian/Alaska Native, non-Hispanic	100	3.79	0.781	98	3.84	0.646	95	3.86	0.845
Two or more races, non-Hispanic	493	3.67	0.847	445	3.74	0.827	407	3.85	0.754
School type									
Public school	10,770	3.69	0.826	9,990	3.73	0.814	9,407	3.80	0.801
Private school	1,112	3.73	0.822	1,012	3.81	0.787	948	3.82	0.818

NOTE: Spring third-grade estimates (X7) are weighted by W7C27P_7T70. Spring fourth-grade estimates (X8) are weighted by W8C28P_8T80. Spring fifth-grade estimates (X9) are weighted by W9C29P_9T90. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to these scales come from the *Temperament in Middle Childhood Questionnaire* (Simonds and Rothbart 2004). The respondent in third grade (round 7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 5. Detail may not sum to totals due to rounding and/or missing data. *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 2014, spring 2015, and spring 2016.

8.1.4 Teacher-Reported Student-Teacher Relationship Scale

The *Student-Teacher Relationship Scale* (*STRS*) (Pianta and Stuhlman 2004) is a 15-item, teacher-reported measure of closeness and conflict between the teacher and child. As part of the spring kindergarten, spring first-grade, spring second-grade, and spring third-grade child-level teacher questionnaire, the teacher was presented with 15 descriptive statements about his or her relationship with the ECLS-K:2011 child and asked to indicate the degree to which each statement applied to their relationship using a 5-point scale ranging from "definitely does not apply" to "definitely applies." The *STRS* was not administered in fourth or fifth grade.

Two scales were developed based on guidelines from the developer of the scale: Closeness and Conflict. The Closeness scale score is the average rating on seven items included in the *STRS*, while the Conflict scale score is the average rating on the other eight items included in 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. The Closeness scale is a measure of the affection, warmth, and open communication that the teacher experiences with the student. The Conflict scale is a measure of the teacher's perception of the negative and conflictual 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 the *STRS* scales are shown in table 8-17.

Table 8-17. Student-Teacher Relationship Scale variable names, descriptions, value ranges, weighted means, and standard deviations for spring kindergarten, spring first grade, spring second grade, and spring third grade: Spring 2011, spring 2012, spring 2013, and spring 2014

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X2CLSNSS	X2 Teacher Report Closeness	15,962	1–5	4.36	0.636
X2CNFLCT	X2 Teacher Report Conflict	15,960	1–5	1.63	0.802
X4CLSNSS	X4 Teacher Report Closeness	13,418	1–5	4.30	0.662
X4CNFLCT	X4 Teacher Report Conflict	13,422	1–5	1.64	0.792
X4KCLSNSS	X4K Teacher Report Closeness	418	1–5	4.27	0.693
X4KCNFLCT	X4K Teacher Report Conflict	418	1–5	1.82	0.875
X6CLSNSS	X6 Teacher Report Closeness	12,681	1–5	4.24	0.687
X6CNFLCT	X6 Teacher Report Conflict	12,683	1–5	1.63	0.794
X7CLSNSS	X7 Teacher Report Closeness	11,894	1–5	4.17	0.712
X7CNFLCT	X7 Teacher Report Conflict	11,901	1–5	1.62	0.782

NOTE: Spring kindergarten estimates (X2) are weighted by W1C0. Spring first-grade estimates (X4) are weighted by W4CS4P_2T0. Spring second-grade estimates (X6) are weighted by W6CS6P_2T0. Spring third-grade estimates (X7) are weighted by W7C27P_7T70. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Stuhlman 2004). The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. Variables that begin with "X4K" are for data collected in the spring first-grade data collection for children who were retained in kindergarten. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.
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, spring 2012, spring 2013, and spring 2014.

Table 8-18 presents the internal consistency reliability coefficients (Cronbach's alpha) for the teacher-reported *STRS* Closeness and Conflict scores for kindergarten, first grade, second grade, and third grade. In the springs of kindergarten, first grade, second grade, and third grade, the Closeness scale (X2CLSNSS, X4CLSNSS, X4KCLSNSS, X6CLSNSS, and X7CLSNSS) has a reliability estimate that

ranges from .86 to .89, and the Conflict scale (X2CNFLCT, X4CNFLCT, X4KCNFLCT, X6CNFLCT, and X7CNFLCT) has a reliability estimate that ranges from .88 to .90.

The study received copyright permission to include item-level data from the *STRS* on the ECLS-K:2011 restricted-use data files. Therefore, these data have been included in the restricted-use data files with the other child-level teacher questionnaire data.[9] Variable names for the item-level data from the spring of kindergarten, the spring of first grade, the spring of second grade, and the spring of third grade begin with "T2," "T4," "T6," and "T7," respectively. Variable names that begin with "T4K" are for item-level data from the spring of first grade for students retained in kindergarten. Because of copyright restrictions, the item-level data do not appear in the public-use data files.

Table 8-18. Reliability estimates for the Student-Teacher Relationship Scale teacher-reported Closeness and Conflict scores for spring kindergarten, spring first grade, spring second grade, and spring third grade: Spring 2011, spring 2012, spring 2013, and spring 2014

Variable name	Description	Number of items	Reliability coefficient
X2CLSNSS	X2 Teacher Report Closeness	7	.89
X2CNFLCT	X2 Teacher Report Conflict	8	.89
X4CLSNSS	X4 Teacher Report Closeness	7	.86
X4CNFLCT	X4 Teacher Report Conflict	8	.89
X4KCLSNSS	X4K Teacher Report Closeness	7	.87
X4KCNFLCT	X4K Teacher Report Conflict	8	.88
X6CLSNSS	X6 Teacher Report Closeness	7	.87
X6CNFLCT	X6 Teacher Report Conflict	8	.90
X7CLSNSS	X7 Teacher Report Closeness	7	.87
X7CNFLCT	X7 Teacher Report Conflict	8	.90

NOTE: Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Stuhlman 2004). The respondent in kindergarten through third grade (rounds 1-7) was the child's classroom teacher. Variables that begin with "X4K" are for data collected in the spring first-grade data collection for children who were retained in kindergarten.
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, spring 2012, spring 2013, and spring 2014.

Means scores for the Closeness and Conflict scores from the *Student-Teacher Relationship Scale* are presented by selected child characteristics in table 8-19 for spring third grade. See the *User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES 2017-285) (Tourangeau et al. 2017); the *ECLS-K:2011 Kindergarten Psychometric Report* (NCES 2018-182) (Najarian et al. 2018a); and the *ECLS-K:2011 First-Grade and Second-Grade*

[9] Item wording is redacted in the questionnaires made available to the public. It is provided in the codebook view of the Electronic Codebook (ECB), restricted version.

Psychometric Report (NCES 2018-183) (Najarian 2018b) for more information on the teacher-reported Closeness and Conflict scores from spring kindergarten, spring first grade, and spring second grade.

Table 8-19. Mean Closeness and Conflict scores from the Student-Teacher Relationship Scale for spring third grade, by child characteristics: Spring 2014

	Spring third grade (X7CLSNSS)			Spring third grade (X7CNFLCT)		
Characteristic	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**11,894**	**4.17**	**0.712**	**11,901**	**1.62**	**0.782**
Sex						
Male	6,086	4.04	0.721	6,092	1.78	0.854
Female	5,793	4.30	0.676	5,794	1.45	0.655
Race/ethnicity						
White, non-Hispanic	5,795	4.25	0.665	5,796	1.56	0.745
Black, non-Hispanic	1,301	4.07	0.725	1,302	1.98	0.971
Hispanic	3,214	4.05	0.764	3,219	1.56	0.726
Asian, non-Hispanic	914	4.07	0.740	914	1.40	0.511
Hawaiian, Other Pacific Islander, non-Hispanic	66	4.10	0.771	66	1.63	0.798
American Indian/Alaska Native, non-Hispanic	100	4.19	0.681	100	1.53	0.721
Two or more races, non-Hispanic	493	4.13	0.723	493	1.68	0.776
School type						
Public school	10,782	4.15	0.715	10,788	1.62	0.787
Private school	1,112	4.32	0.658	1,113	1.57	0.711

NOTE: Spring third-grade estimates (X7) are weighted by W7C27P_7T70. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to these scales come from the *Student-Teacher Relationship Scale* (Pianta and Stuhlman 2004). Used with permission. The respondent in third grade (round 7) was the child's classroom teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 5. Detail may not sum to totals due to rounding and/or missing data. *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 2014.

8.1.5 Teacher-Reported Working Memory

In third grade, teachers reported on child behaviors related to working memory. In third and fourth grades, parents also reported on child behaviors related to working memory, using items that were identical to the teacher-reported items. The third- and fourth-grade rounds were the only rounds of data collection in which parent- or teacher-reported working memory was obtained.

In the spring of third grade, teachers and parents reported on child behaviors related to working memory. Teachers were asked the four items as part of the child-level teacher questionnaire, and parents

were asked the same four items in the parent interview.[10] The items are four of the ten items that make up the Parent and Teacher Forms of the Working Memory Scale of the Behavior Rating Inventory of Executive Function (BRIEF).[11] Items from the BRIEF Working Memory Scale measure "the capacity to hold information in mind for the purpose of completing a task" (Gioia et al. 2000, p. 19). Teachers and parents were presented with statements that describe child behaviors related to working memory, and they were asked to rate how often (Never, Sometimes, Often) the child has had problems with these behaviors over the past 6 months. See section 8.2.1 for specific information about the working memory items obtained from parents.

Item-level data on working memory obtained from the teacher in third grade are provided on the data files. Variables from the spring third-grade child-level teacher questionnaire begin with "T7."

The data file also contains a scale score for teacher-reported working memory (X7TWKMEM). A scale score was computed when the respondent provided a rating on at least three of the four 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 indicate that the child exhibited more behaviors indicating problems with working memory. That is, higher scores indicate *worse* working memory. Lower scale scores indicate fewer difficulties related to working memory, and, therefore, indicate *better* working memory. The variable names, descriptions, value ranges, weighted means, and standard deviations for these scales are shown in table 8-20.

[10] The items used for teachers and parents were the same and matched the items from the Behavior Rating Inventory of Executive Function (BRIEF). The instructions were adapted from the instructions on the cover of the BRIEF questionnaire to be appropriate for the mode of data collection used in this study. The instructions were adapted to be as similar as possible to the intent of the BRIEF instructions. The instructions varied slightly for parents and teachers because the parent items were administered within the parent interview and teachers completed items in a hard-copy questionnaire.

[11] The ECLS-K:2011 used four of ten items from the Teacher Form of the BRIEF and 4 of 10 items from the Parent Form of the BRIEF. The items used were adapted and reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Behavior Rating Inventory of Executive Function by Gerard A. Gioia, Peter K. Isquith, Steven C. Guy, and Lauren Kenworthy, Copyright 1996, 1998, 2000 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.

Table 8-20. Teacher-reported Working Memory variable names, descriptions, value ranges, weighted means, and standard deviations for spring third grade: Spring 2014

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X7TWKMEM	X7 Teacher Report Working Memory	11,870	1-3	1.58	0.607

NOTE: Spring third-grade estimates (X7) for teacher-reported data are weighted by W7C27P_7T70. Items contributing to this scale come from the Working Memory Scale of the Behavior Rating Inventory of Executive Function (BRIEF) Teacher Form. The items used were adapted and reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Behavior Rating Inventory of Executive Function by Gerard A. Gioia, Peter K. Isquith, Steven C. Guy, and Lauren Kenworthy, Copyright 1996, 1998, 2000 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc. The respondent in third grade (round 7) was the child's classroom teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014.

Table 8-21 presents internal consistency reliability coefficients (Cronbach's alpha) for the teacher Working Memory scale. The teacher Working Memory scale for spring third grade has an internal consistency reliability coefficient of .91.

Table 8-21. Reliability estimates for the teacher-reported Working Memory scores for spring third grade: Spring 2014

Variable name	Description	Number of items	Reliability coefficient
X7TWKMEM	X7 Teacher Report Working Memory	4	.91

NOTE: Items contributing to these scales come from the Working Memory Scale of the Behavior Rating Inventory of Executive Function (BRIEF) Teacher Form. The items used were adapted and reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Behavior Rating Inventory of Executive Function by Gerard A. Gioia, Peter K. Isquith, Steven C. Guy, and Lauren Kenworthy, Copyright 1996, 1998, 2000 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc. The respondent in third grade (round 7) was the child's classroom teacher.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014.

Mean scores for the spring third-grade teacher-reported Working Memory scale are presented by child characteristics in table 8-22.

Table 8-22. Mean teacher-reported Working Memory scale for spring third grade, by child characteristics: Spring 2014

Characteristic	Spring third grade (X7TWKMEM) Number	Mean	*SD*
Total sample	**11,870**	**1.58**	**0.607**
Sex			
Male	6,081	1.70	0.631
Female	5,775	1.45	0.550
Race/ethnicity			
White, non-Hispanic	5,786	1.55	0.600
Black, non-Hispanic	1,299	1.75	0.636
Hispanic	3,202	1.57	0.596
Asian, non-Hispanic	912	1.42	0.537
Hawaiian, Other Pacific Islander, non-Hispanic	67	1.45	0.573
American Indian/Alaska Native, non-Hispanic	100	1.56	0.558
Two or more races, non-Hispanic	493	1.60	0.625
School type			
Public school	10,761	1.58	0.609
Private school	1,109	1.54	0.580

NOTE: Spring third-grade estimates (X7) are weighted by W7C27P_7T70. See the ECLS-K:2011 data file User's Manuals for explanation of the weighting variables. Items contributing to this scale come from the Working Memory Scale of the Behavior Rating Inventory of Executive Function (BRIEF) Teacher Form and Parent Form. The items used were adapted and reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Behavior Rating Inventory of Executive Function by Gerard A. Gioia, Peter K. Isquith, Steven C. Guy, and Lauren Kenworthy, Copyright 1996, 1998, 2000 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc. The respondent in third grade (round 7) was the child's classroom teacher. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 3. Detail may not sum to totals due to rounding and/or missing data. *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 2014.

8.1.6 Teacher-Reported Measures of the Peer Context and the Child's Relationships with Peers

Teachers reported their perceptions of the child's peer relationships in the child-level teacher questionnaire in spring of second grade and spring of third grade and in the reading subject-specific child-level teacher questionnaire in spring of fourth grade and spring of fifth grade. There are questions in the parent interview that complement the teacher-reported information on peer relationships. See section 8.2.2 for information about parent-reported items about the child's peer relationships and friends.

Exhibit 8-1 shows the constructs on peer relationships included in the second-, third-, fourth-, and fifth-grade child-level teacher questionnaires and the corresponding item-level variables along with

their sources. In second, third, fourth, and fifth grades, teachers provided information on peer victimization, both with the child as the victim and with the child as the aggressor. In the spring of third grade, spring of fourth grade, and spring of fifth grade, teachers were asked about whether the child was excluded or ignored by peers and about whether the child exhibited prosocial behaviors with peers. In the spring of fourth grade and the spring of fifth grade, teachers were asked about the behaviors of the peers in the child's peer group and about the child's social skills with peers. These items were adapted from existing scales and were used with the permission of the authors. Data for the individual items are included in the K–5 data file. Variable names for the item-level data from the child-level teacher questionnaire in the spring of second grade and the spring of third grade begin with "T6" and "T7," respectively. Variable names from the item-level data from the reading subject-specific child-level teacher questionnaire for the spring of fourth grade and the spring of fifth grade begin with "G8" and "G9," respectively. Composite variables for each construct are not provided; it is left to analysts to decide how best to use these data in their analyses.

It was beyond the scope of ECLS-K:2011 to create a composite variable for each construct listed in exhibit 8-1. The fact that no construct was computed and provided on the dataset is not an indication that the items did not represent the construct or that there was lack of evidence for the construct. In some instances, only a subset of items was selected from an existing scale and wording was adapted to fit into questionnaires administered. In some instances, there are multiple ways in which the items could be used, and it is left to the researcher's discretion how to best use the items. Sources for the items are provided in exhibit 8-1. Data users are encouraged to examine how other researchers have used these items.

Although composites were not computed for these items, preliminary analyses were conducted to examine scale reliability. Cronbach's alpha was examined for each set of items as a measure of internal consistency, or how closely related the set of items are as a group.

Exhibit 8-1. Teacher-reported item-level variables on peer relationships in spring second grade, spring third grade, spring fourth grade, and spring fifth grade: Spring 2013, spring 2014, spring 2015, and spring 2016

Construct/scale	Grade administered	Number of items	Item-level variable names
Peer Victimization (child as victim)[1]	2-5	4	T6OSTEAS/T7OSTEAS/G8OSTEAS/G9OSTEAS; T6OSLIES/T7OSLIES/G8OSLIES/G9OSLIES; T6OSPUSH/T7OSPUSH/G8OSPUSH/G9OSPUSH; T6OSLFTO/T7OSLFTO/G8OSLFTO/G9OSLFTO
Peer Victimization (child as aggressor)[1]	2-5	4	T6TSTEAS/T7TSTEAS/G8TSTEAS/G9TSTEAS; T6TSLIES/T7TSLIES/G8TSLIES/G9TSLIES; T6TSPUSH/T7TSPUSH/G8TSPUSH/G9TSPUSH; T6TSLFTO/T7TSLFTO/G8TSLFTO/G9TSLFTO
Excluded by Peers[2]	3-5	4	T7PLYMTE/G8PLYMTE/G9PLYMTE; T7PAVOID/G8PAVOID/G9PAVOID; T7EXLUED/G8EXLUED/G9EXLUED; T7IGNRED/G8IGNRED/G9IGNRED
Prosocial with Peers[2]	3-5	5	T7OTDIST/G8OTDIST/G9OTDIST; T7ISKIND/G8ISKIND/G9ISKIND; T7COPRTV/G8COPRTV/G9COPRTV; T7CNMORL/G8CNMORL/G9CNMORL; T7HLPUPS/G8HLPUPS/G9HLPUPS
Positive Peer Group[3]	4-5	9	G8GOODGP/G9GOODGP; G8WORYGP/G9WORYGP; G8BADINF/G9BADINF; G8SUPVIS/G9SUPVIS; G8TRBLGP/G9TRBLGP; G8EXCSTU/G9EXCSTU; G8HRDWKR/G9HRDWKR; G8FUNGRP/G9FUNGRP; G8KINDGP/GKINDGP
Social Skills with Peers[4]	4-5	4	G8UNDFEL/G9UNDFEL; G8INTPER/G9INTPER, G8SOLINT/G9SOLINT, G8EFFBEV/G9EFFBEV

[1] Peer victimization items were adapted from a 21-item scale. Espelage, D.L., and Holt, M. (2001). Bullying and Peer Victimization During Early Adolescence: Peer Influences and Psychosocial Correlates. *Journal of Emotional Abuse, 2:* 123–142.
[2] Adapted from the Child Behavior Scale © 2010 Gary W. Ladd. Used with permission. A subset of items from the Excluded by Peers and Prosocial with Peers scales from the Child Behavior Scale were adapted and used.
[3] Adapted from Vandell, D.L. (2001). *Relationships With Peers: Part D (Teacher).* Unpublished scale, National Institute of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development, Form FSV10G3. These items reflect positive and negative peer group characteristics. The NICHD Study of Early Child Care and Youth Development decided to form one composite for "Positive Peer Group" with these items, reverse coding four of the nine items when creating a composite.
[4] Adapted from Pierce, K.M., Hamm, J.V., and Vandell, D.L. (1999). Experiences in After-school Programs and Children's Adjustment in First-grade Classrooms. *Child Development, 70*, 756-767. These items include four of seven items from the "Mock Report Card" (e.g., Form FSV08G3) used in the NICHD Study of Early Child Care and Youth Development and were originally adapted from Coie and Dodge (1988).
NOTE: The respondent in second grade and third grade (rounds 6 and 7) was the child's classroom teacher. The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher.
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, spring 2014, spring 2015, and spring 2016.

Peer victimization. In spring of second, third, fourth, and fifth grades, teachers were asked about peer victimization, with separate questions about the child as a victim and the child as the aggressor.

The questions that ask about child as a victim were also asked of parents in the spring of second grade and the spring of third grade.[12] Each set of four questions asked about different types of peer victimization, and researchers may choose not to combine them or to combine some or all of them, depending upon the research question. In the spring of second grade, the internal consistency reliability coefficient (Cronbach's alpha) is .84 for the four peer victimization items with the child as victim, and .90 for the four peer victimization items with the child as the aggressor. In the spring for third grade, the internal consistency is .82 for the peer victimization items with the child as victim, and .89 for the peer victimization items with the child as the aggressor. In the spring of fourth grade, the internal consistency is .83 for the peer victimization items with the child as victim, and .90 for the peer victimization items with the child as the aggressor. Fifth-grade internal consistency is .84 for the items with the child as the victim and .90 for the items with the child as the aggressor.

Excluded by peers and prosocial with peers. In spring third, fourth, and fifth grades, teachers were asked about whether the child was excluded or ignored by peers and whether the child demonstrated prosocial behaviors toward peers. The four excluded by peers items and the five prosocial with peers items were selected from a larger set of items from the Child Behavior Scale (Ladd 2010). In spring third grade, the internal consistency reliability coefficient is .89 for the excluded by peers items and .85 for the prosocial with peers items. In spring fourth grade, the internal consistency is .90 for the excluded by peers items and .86 for the prosocial with peers items. In spring fifth grade, the excluded by peers items have an internal consistency of .90 and the prosocial with peers items have an internal consistency of .86.

Positive peer group and negative peer group. Teachers rated nine items about the child's peer group in fourth and fifth grades. Other studies have combined these nine items into a scale of the positive peer group by reverse coding the four items that are negatively worded (see footnote in exhibit 8-1). However, it is possible to consider two separate scales, one for positive characteristics of the peer group and one for negative characteristics of the peer group. In fourth grade, internal consistency of the nine items (five positive items and four negative items reverse coded) is .93. Looking separately at the positive peer group items and the negative peer group items in spring of fourth grade, the internal consistency of the five positive items is .90 and the internal consistency of the four negative items is .91. In spring of fifth grade, the internal consistency of the nine items is .94, and the internal consistency of the five positive items and the four negative items is .90 and .92, respectively.

[12] In the spring of second grade, parents were only asked about three of the four types of peer victimization that teachers were asked about. Parents were not asked whether other children ever told lies or untrue stories about their child. A question about this type of victimization was added to the parent interview in third grade so that teachers and parents would be asked about the same types of victimization.

Social skills with peers. In fourth and fifth grades, teachers were asked four items about the child's social skills with peers. Teachers were asked about whether the child is good at understanding the feelings of others, accurately interprets a peer's intentions, generates good quality solutions to interpersonal problems, and is aware of the effects of his or her behavior on others. The internal consistency reliability coefficient is .95 for the spring of fourth grade and .95 for the spring of fifth grade.

8.1.7 Teacher-Reported Perceptions of Child's School Liking

In the spring of fourth grade and the spring of fifth grade, teachers and parents reported their perceptions of the child's school liking and avoidance behaviors using items adapted from the parent and teacher versions of the School Liking and Avoidance Questionnaire (SLAQ) (Ladd and Price 1987; Ladd 1990). Teachers rated perceptions of school liking with seven items, four positively worded items (e.g., "Likes to come to school") and three negatively worded items (e.g., "Dislikes school"), on a 3-point scale to indicate whether the item "doesn't apply (seldom displays this behavior)," "sometimes applies (occasionally displays this behavior)," or "certainly applies (often displays this behavior)." Ladd used these seven items to create a single teacher-reported school liking construct by combining these seven items (reverse scoring the negatively worded items). Exhibit 8-2 presents the item-level school liking variables. See section 8.2.3 for more information on the items rated by parents on the parent's perception of school avoidance behaviors. Composite variables for these teacher constructs are not provided; it is left to analysts to decide how best to use these data in their analyses.

Exhibit 8-2. Teacher-reported item-level variables on school liking in spring fourth grade and spring fifth grade: Spring 2015 and spring 2016

Construct/scale	Grade administered	Number of items	Item-level variable names
Teacher-report School Liking[1]	4–5	7	G*LIKSCH, G*DISLSH, G*FUNSCH, G*LBESCH, G*UNHAPY, G*ENJACT, G*GRNACT

[1] Adapted from the teacher version of the School Liking and Avoidance Questionnaire (SLAQ; Adapted from Ladd and Price, 1987; Ladd, 1990)
NOTE: The respondent in fourth grade (round 8) and fifth grade (round 9) was the child's reading and language arts teacher. An asterisk "*" is a placeholder for round number in variable names. Fourth grade is round 8, and fifth grade is round 9. For example, the variable G*LIKSCH is listed in the table; this indicates that the variables G8LIKSCH and G9LIKSCH are available in the dataset.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2015 and spring 2016.

Although composites were not computed for these items, preliminary analyses were conducted to examine scale reliability. Cronbach's alpha was examined for the teacher-reported school liking variables as a measure of internal consistency, or how closely related the set of items are as a group. The internal

consistency of all seven items was examined. Teachers were asked to rate how often each of the following items applies to the study child: "Likes to come to school," "Dislikes school," "Has fun at school," "Likes being in school," "Seems unhappy in school," "Enjoys most classroom activities," "Groans or complains about suggested activities." The internal consistency reliability of these 7 items, with the negatively worded items reverse coded, was .89 for the spring of fourth grade and for the spring of fifth grade.

The internal consistency was also examined for the 4 positively worded items and separately for the three negatively worded items. The four positively worded school liking items had an internal consistency reliability coefficient of .90 in both the spring of fourth grade and the spring of fifth grade. The three negatively worded items about disliking school, being unhappy at school, and groaning or complaining about activities at school had an internal consistency reliability coefficient of .74 in both the spring of fourth grade and the spring of fifth grade.

8.2 Parent Measures

This section presents parent-reported perceptions of the child's skills, behaviors, and relationships in third-grade, fourth-grade, and fifth-grade data collections. This section provides information on the parent's perception of the child's working memory, peer relationships and friendships, and school avoidance behaviors. In the fall and spring of kindergarten and in the spring of first grade, parents were asked about learning behaviors (approaches to learning) and social skills. For information about these indirect measures, see the *User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES 2017-285) (Tourangeau et al. 2017); the *ECLS-K:2011 Kindergarten Psychometric Report* (NCES 2018-182) (Najarian 2018a); and the *ECLS-K:2011 First-Grade and Second-Grade Psychometric Report* (NCES 2018-183) (Najarian 2018b).

8.2.1 Parent-Reported Working Memory

In third and fourth grades, parents reported on child behaviors related to working memory. In third grade, teachers also reported on child behaviors related to working memory. The third- and fourth-grade rounds were the only rounds of data collection in which parent- or teacher-reported working memory was obtained.

Parents were asked four items as part of the parent interview in third and fourth grades.[13] Teachers were asked the same four items in the spring of third grade.[14] The items are four of the ten items that make up the Parent Form of the Working Memory Scale of the Behavior Rating Inventory of Executive Function (BRIEF).[15] Items from the BRIEF Working Memory Scale measure "the capacity to hold information in mind for the purpose of completing a task" (Gioia et al. 2000, p. 19). Parents were presented with statements that describe child behaviors related to working memory, and they were asked to rate how often (Never, Sometimes, Often) the child has had problems with these behaviors over the past 6 months. See section 8.1.5 for specific information about the working memory items obtained from teachers.

Item-level data are provided on the data files. Variables for the item-level data from the spring third-grade and spring fourth-grade parent interviews begin with "P7" and "P8," respectively.

The data file also contains scale scores for parent-reported working memory in third and fourth grades (X7PWKMEM, X8PWKMEM). For the parent scale score, a score was computed when the respondent provided a rating on at least three of the four 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 indicate that the child exhibited more behaviors indicating problems with working memory. That is, higher scores indicate *worse* working memory. Lower scale scores indicate fewer difficulties related to working memory, and, therefore, indicate *better* working memory. The variable names, descriptions, value ranges, weighted means, and standard deviations for these scales are shown in table 8-23.

Table 8-24 presents internal consistency reliability coefficients (Cronbach's alpha) for the parent Working Memory scale. The parent Working Memory scale has an internal consistency reliability coefficient of .81 in the spring of third grade and .83 in the spring of fourth grade.

[13] The items rated by parents were the same each round and matched the items from the Behavior Rating Inventory of Executive Function (BRIEF). The instructions were adapted from the instructions on the cover of the BRIEF questionnaire to be appropriate for the mode of data collection used in this study. The instructions were adapted to be as similar as possible to the intent of the BRIEF instructions.

[14] The items used for teachers and parents were the same and matched the items from the Behavior Rating Inventory of Executive Function (BRIEF). The instructions were adapted from the instructions on the cover of the BRIEF questionnaire to be appropriate for the mode of data collection used in this study. The instructions were adapted to be as similar as possible to the intent of the BRIEF instructions. The instructions varied slightly for parents and teachers because the parent items were administered within the parent interview and teachers completed items in a hard-copy questionnaire.

[15] The ECLS-K:2011 used four of ten items from the Parent Form of the BRIEF. The items used were adapted and reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Behavior Rating Inventory of Executive Function by Gerard A. Gioia, Peter K. Isquith, Steven C. Guy, and Lauren Kenworthy, Copyright 1996, 1998, 2000 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.

Table 8-23. Parent-reported Working Memory variable names, descriptions, value ranges, weighted means, and standard deviations for spring third grade and spring fourth grade: Spring 2014 and spring 2015

Variable name	Description	*n*	Range of possible values	Weighted mean	Standard deviation
X7PWKMEM	X7 Parent Report Working Memory	10,689	1-3	1.70	0.535
X8PWKMEM	X8 Parent Report Working Memory	10,297	1-3	1.71	0.547

NOTE: Spring third-grade estimates (X7) for parent-reported data are weighted by W7C27P_7A0. Spring fourth-grade estimates (X8) for parent-reported data are weighted by W8C28P_8A0. Items contributing to these scales come from the Working Memory Scale of the Behavior Rating Inventory of Executive Function (BRIEF) Parent Form. The items used were adapted and reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Behavior Rating Inventory of Executive Function by Gerard A. Gioia, Peter K. Isquith, Steven C. Guy, and Lauren Kenworthy, Copyright 1996, 1998, 2000 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014 and spring 2015.

Table 8-24. Reliability estimates for the parent-reported Working Memory scores for spring third grade and spring fourth grade: Spring 2014 and spring 2015

Variable name	Description	Number of items	Reliability coefficient
X7PWKMEM	X7 Parent Report Working Memory	4	.81
X8PWKMEM	X8 Parent Report Working Memory	4	.83

NOTE: Items contributing to these scales come from the Working Memory Scale of the Behavior Rating Inventory of Executive Function (BRIEF) Parent Form. The items used were adapted and reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Behavior Rating Inventory of Executive Function by Gerard A. Gioia, Peter K. Isquith, Steven C. Guy, and Lauren Kenworthy, Copyright 1996, 1998, 2000 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2014 and spring 2015.

Mean scores for the parent-reported Working Memory scale in spring third grade and spring fourth grade are presented by child characteristics in table 8-25.

Table 8-25. Mean parent-reported Working Memory scale for spring third grade and spring fourth grade, by child characteristics: Spring 2014 and spring 2015

Characteristic	Spring third grade (X7PWKMEM)			Spring fourth grade (X8PWKMEM)		
	Number	Mean	*SD*	Number	Mean	*SD*
Total sample	**10,689**	**1.70**	**0.535**	**10,297**	**1.71**	**0.547**
Sex						
Male	5,458	1.78	0.548	5,239	1.79	0.558
Female	5,231	1.63	0.509	5,058	1.63	0.523
Race/ethnicity						
White, non-Hispanic	5,418	1.73	0.541	5,211	1.75	0.557
Black, non-Hispanic	1,054	1.73	0.537	975	1.74	0.530
Hispanic	2,731	1.64	0.523	2,687	1.63	0.535
Asian, non-Hispanic	863	1.57	0.456	824	1.58	0.475
Hawaiian, Other Pacific Islander, non-Hispanic	53	1.59	0.483	47	1.47	0.432
American Indian/Alaska Native, non-Hispanic	85	1.66	0.548	88	1.67	0.519
Two or more races, non-Hispanic	485	1.79	0.548	465	1.78	0.542
School type						
Public school	9,154	1.71	0.538	8,707	1.71	0.548
Private school	984	1.68	0.506	904	1.69	0.526

NOTE: Spring third-grade estimates (X7) for parent-reported data are weighted by W7C27P_7A0. Spring fourth-grade estimates (X8) for parent-reported data are weighted by W8C28P_8A0. Items contributing to these scales come from the Working Memory Scale of the Behavior Rating Inventory of Executive Function (BRIEF) Parent Form. The items used were adapted and reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Behavior Rating Inventory of Executive Function by Gerard A. Gioia, Peter K. Isquith, Steven C. Guy, and Lauren Kenworthy, Copyright 1996, 1998, 2000 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc. The unweighted sample *n* indicates the number of cases with valid data regardless of the presence of a valid analytic weight. The range of possible values is 1 to 3. Detail may not sum to totals due to rounding and/or missing data. *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 2014 and spring 2015.

8.2.2 Parent-Reported Measures of the Child's Friendships and Relationships with Peers

Parents reported their perceptions of the child's peer relationships in the parent interview in the spring of second, third, fourth, and fifth grades. Questions in the parent interviews complement the teacher-reported information on peer relationships. See section 8.1.6 for information on teacher-reported measures of the peer context and the child's relationship with peers.

Exhibit 8-3 shows the constructs on peer relationships included in the second-, third-, fourth-, and fifth-grade parent interviews and the corresponding item-level variables along with their sources. Parents provided information on peer victimization (child as victim) in second and third grades. The teacher- and parent-provided information complements information collected from children on peer

victimization, which is described above in chapter 7. In fourth and fifth grades, parents were asked about their child's friendships. Parents were asked how many close friends their child has and what type of influence their child's best friend has on their child. Data for the individual items are included on the data files. Variable names for the item-level data from the parent interviews in the spring of second, third, fourth, and fifth grades begin with "P6," "P7," "P8," and "P9," respectively.

Exhibit 8-3. Parent-reported item-level variables on peer relationships and friendships in spring second grade, spring third grade, spring fourth grade, and spring fifth grade: Spring 2013, spring 2014, spring 2015, and spring 2016

Construct/scale	Number of items (grade)	Response categories	Item-level variable names
Peer Victimization [1] (child as victim)	3 (second grade) 4 (third grade)	Yes, No	P*OTHTEA P7OTHLIE[2] P*OTHHIT P*OTHEXC
Peer Victimization [1] (child as victim)	3 (second grade) 4 (third grade)	Rarely, Sometimes, Often, Very Often	P*OFTTEA P7OFTLIE[2] P*OFTHIT P*OFTEXC
Number of Close Friends	1 (fourth/fifth grade)	Number	P*NUMFRD
Influence of Best Friend	1 (fourth/fifth grade)	Always a good influence, Usually a good influence, Neither a good nor a bad influence, Usually a bad influence, Always a bad influence	P*FRINFL

[1] Peer victimization items were adapted from a 21-item scale. Espelage, D.L. and Holt, M. (2001). Bullying and Peer Victimization During Early Adolescence: Peer Influences and Psychosocial Correlates. *Journal of Emotional Abuse, 2:* 123–142.
[2] In second grade, parents were not asked about whether other children told lies or untrue stories about their child. An item was added in third grade so that parents, teachers, and children were asked about the same forms of peer victimization.
NOTE: An asterisk "*" is a placeholder for round number in variable names. Third grade is round 7, fourth grade is round 8, and fifth grade is round 9. For example, the variable P*OTHTEA is listed in the table; this indicates that the variables P7OTHTEA and P8OTHTEA are available in the dataset.
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, spring 2014, spring 2015, and spring 2016.

Parents were asked about three or four types of peer victimization. In second grade, parents were asked about 3 kinds of peer victimization, but in third grade a fourth type of peer victimization was added to the parent interview to match what was asked in the teacher and child questionnaires. Parents were asked about each type of peer victimization using a two-part question. Parents were first asked whether a particular type of peer victimization occurred and, if so, they were asked how often it occurred. For example, parents were asked "During this school year have other children ever teased, made fun of, or called {CHILD} names?" If they answered "yes," then parents were asked "How often has this happened? Would you say…" and parent were asked to rate the frequency as "rarely," "sometimes," "often," or "very

often?" If they answered "no," then parents were asked about the next type of peer victimization. In second grade, parents were also asked "During this school year have other children ever pushed, shoved, slapped, hit, or kicked {CHILD}?" and "During this school year have other children ever intentionally excluded or left {CHILD} out from playing with them?" In third grade, parents were asked one additional question, which was asked following the question about teasing/name calling and before the question about physical aggression toward the child: "During this school, have other children ever told lies or untrue stories about {CHILD}?" Two-part questions were used for parents because of the mode of data collection. Two-part questions were easier to administer in a telephone interview, and they took less time to administer when the child was not a victim. The parent variables associated with each part of the two-part question can be combined to create variables with the same five response categories (Never, Rarely, Sometimes, Often, Very often) used on the teacher and child questionnaires.

Composite variables for the peer victimization items and friendship items are not provided; it is left to analysts to decide how best to use these data in their analyses. These questions ask about different types of victimization, and researchers may choose not to combine them or to combine some or all of them, depending upon the research question. This is left to the discretion of the data user. However, preliminary analyses were conducted to examine the internal consistency reliability of these items (Cronbach's alpha).[16] The internal consistency of the three parent-reported items in second grade is .72, and the internal consistency of the four parent-reported items in third grade is .78.

8.2.3 Parent-Reported School Avoidance

In the spring of fourth grade and the spring of fifth grade, parents and teachers reported their perceptions of the child's school liking and avoidance behaviors using items adapted from the parent and teacher versions of the School Liking and Avoidance Questionnaire (SLAQ) (Ladd and Price 1987; Ladd 1990). Parents rated five items about the parent's perception of their child's school avoidance behaviors on a 5-point scale, using response items similar to the SLAQ (Almost never, Rarely, Sometimes, A lot, Almost always). Ladd used these five items to create a single parent-reported school avoidance scale (exhibit 8-4). Composite variables for the parent constructs are not provided; it is left to analysts to decide how best to use these data in their analyses. See section 8.1.7 for information on items on the teacher's perception of the child's school liking behaviors.

[16] Because the parent items were asked in a two-part question, the two parts were combined to create a single variable that was used to examine internal consistency. This was done so that the parent items would be constructed like the one-part teacher- and child-reported items. Specifically, if parents responded "no" to whether the particular type of peer victimization occurred, this was assigned as a "never" rating. The combined parent variable had 5 rating points that were identical to the one-part question asked of teachers and children: 1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = very often.

Although composites were not computed for these parent items, preliminary analyses were conducted to examine scale reliability. Cronbach's alpha was examined for the parent-reported school avoidance variables as a measure of internal consistency. Parents were asked to rate how often the child: "Makes up reasons to stay home from school," "Seems to dread going to school," "Becomes upset when it's time to go to school in the morning," "Asks to stay home from school," and "Complains about going to school." The internal consistency reliability of these five items was .87 for the spring of fourth grade and .88 for the spring of fifth grade.

Exhibit 8-4. Parent-reported item-level variables on school avoidance in spring fourth grade and spring fifth grade: Spring 2015 and spring 2016

Construct/scale	Grade administered	Number of items	Item-level variable names
Parent-reported School Avoidance[1]	4-5	5	P*MKREAS, P*CDREAD, P*CUPSET, P*STAYHM, P*CMPLNS

[1] Adapted from the parent version of the School Liking and Avoidance Questionnaire (SLAQ; Adapted from Ladd and Price, 1987; Ladd, 1990)
NOTE: An asterisk "*" is a placeholder for round number in variable names. Fourth grade is round 8, and fifth grade is round 9. For example, the variable P*MKREAS is listed in the table; this indicates that the variables P8MKREAS and P9MKREAS are available in the dataset.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), spring 2015 and spring 2016.

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.

Berndt, T.J., and Keefe, K. (1995). Friends' Influence on Adolescents' Adjustment to School. *Child Development*, *66:* 1312-1329.

Bock, R.D., Thissen, D., and Zimowski, M.F. (1997). IRT Estimation of Domain Scores. *Journal of Educational Measurement, 34*: 197-211.

Brunton-Smith, I., Sturgis, P., and Leckie, G. (2017). Detecting and Understanding Interviewer Effects on Survey Data by Using a Cross-Classified Mixed Effects Location–Scale Model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(2): 551-568.

Coie, J.D., and Dodge, K.A. (1988). Multiple Sources of Data on Social Behavior and Social Status in School: A Cross-Age Comparison. *Child Development, 59:* 815-829.

Cole, N.S., and Moss, P.A. (1989). Bias in Test Use. In R.L. Linn (Ed.), *The American Council on Education/Macmillan Series on Higher Education. Educational Measurement* (pp. 201-219). New York, NY, England: Macmillan Publishing Co, Inc; American Council on Education.

Crick, N.R., and Grotpeter, J.K. (1995). Relational Aggression, Gender, and Social Psychological Adjustment. *Child Development*, *66:* 710-722.

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York, NY: Holt, Rinehart, and Winston.

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 Holland, P.W. (1992). *DIF Detection and Description: Mantel–Haenszel and Standardization* (Research Report No. RR-92-10). Princeton, NJ: Educational Testing Service.

Dorans, N.J., and Kulick, E. (2006). Different Item Functioning on the Mini-Mental State Examination: An Application of the Mantel-Haenszel and Standardization Procedures. *Medical Care*, *44*(11) Suppl3: S107-S114.

Duckworth, A.L., and Quinn, P.D. (2009). Development and Validation of the Short Grit Scale (Grit–S). *Journal of Personality Assessment, 91*:2, 166-174.

Eriksen, B.A., and Eriksen, C.W. (1974). Effects of Noise Letters Upon the Identification of a Target Letter in a Nonsearch Task. *Perception and Psychophysics, 16:* 143-149.

Espelage, D.L., and Holt, M. (2001). Bullying and Peer Victimization During Early Adolescence: Peer Influences and Psychosocial Correlates. *Journal of Emotional Abuse*, *2*: 123-142

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

Gioia, G.A., Isquith, P.K., Guy, S.C., and Kenworthy, L. (2000). Behavior Rating Inventory of Executive Function. Lutz, FL: Psychological Assessment Resources.

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

Haberman, S.J. (2013). *A General Program for Item-Response Analysis That Employs the Stabilized Newton-Raphson Algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing 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.

Holcomb, R., Williams, R.C., and Richards, P.S. (1985). The Elements of Attachment: Relationship Maintenance and Intimacy. *Journal of the Delta Society, 2*(1): 28-34.

Holland, P.W., and Thayer, D.T. (1986). *Differential Item Functioning and the Mantel-Haenszel Procedure* (ETS Research Report No. 86-31). Princeton, NJ: Educational Testing Service.

Hox, J.J. (1994). Hierarchical Regression Models for Interviewer and Respondent Effects. *Sociological Methods & Research*, *22*(3): 300-318.

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 https://www.literacyworldwide.org/docs/default-source/where-we-stand/phonemic-awareness-position-statement.pdf?sfvrsn=944ea18e_6.

International Reading Association and the National Association for the Education of Young Children. (1998). *Learning to Read and Write: Developmentally Appropriate Practices for Young Children: Continuum of Children's Development in Early Reading and Writing.* A joint position statement of the International Reading Association and the National Association for the Education of Young Children. *Young Children*, *53*(4): 30-46. Retrieved June 20, 2012 from http://www.naeyc.org/files/naeyc/file/positions/PSREAD98.PDF.

Kerr, M., and Stattin, H. (2000). What Parents Know, How They Know It, and Several Forms of Adolescent Adjustment: Further Support for a Reinterpretation of Monitoring. *Developmental Psychology, 36:* 366-380.

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). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Kolen, M.J., and Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (Statistics for Social and Behavioral Sciences)*. (2nd ed.). New York: Springer.

Ladd, G.W. (1990). Having Friends, Keeping Friends, Making Friends and Being Liked by Peers in the Classroom: Predictors of Children's Early School Adjustment? *Child Development, 61:* 1081-1100.

Ladd, G.W. (2010). The Child Behavior Scale: Applications and Research Findings. Mesa, AZ: Parkview Publications.

Ladd, G.W., Kochendorfer, B.J., and Coleman, C.C. (1996). Friendship Quality as a Predictor of Young Children's Early School Adjustment. *Child Development*, *67:* 1103-1118.

Ladd, G.W., and Price, J. (1987). Predicting Children's Social and School Adjustment Following the Transition from Preschool to Kindergarten. *Child Development*, *58*(5): 1168-1189.

La Greca, A.M., and Stone, W.L. (1993). Social Anxiety Scale for Children—Revised: Factor Structure and Concurrent Validity. *Journal of Clinical Child Psychology*, *22*(1): 17-27.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Publishers.

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

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.

Marsh, H.W. (1992). *Self Description Questionnaire (SDQ) I: A Theoretical and Empirical Basis for the Measurement of Multiple Dimensions of Preadolescent Self-Concept. An Interim Test Manual and Research Monograph.* Macarthur, New South Wales, Australia: University of Western Sydney, Faculty of Education.

McGrew, K.S., and Woodcock, R.W. (2001). Technical Manual. *Woodcock-Johnson III*. 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 Software.

Mislevy, R.J., Johnson, E.G., and Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*: 131-154.

Muraki, E.J., and Bock, R.D. (1987). *BIMAIN: A Program for Item Pool Maintenance in the Presence of Item Parameter Drift and Item Bias.* Mooresville, IN: Scientific Software.

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

Najarian, M., Tourangeau, K., Nord, C., and Wallner-Allen, K. (2018a). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), Kindergarten Psychometric Report* (NCES 2018-182). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Najarian, M., Tourangeau, K., Nord, C., and Wallner-Allen, K. (2018b). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First-Grade and Second-Grade Psychometric Report* (NCES 2018-183). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

National Assessment Governing Board. (2000). *Reading Framework for the National Assessment of Educational Progress: 1992–2000*. Washington, DC: Author.

National Assessment Governing Board. (2008). *Reading Framework for the 2009 National Assessment of Educational Progress.* Washington, DC: U.S. Government Printing Office.

National Assessment Governing Board. (2010). *Reading Framework for the 2011 National Assessment of Educational Progress.* Washington, DC: U.S. Government Printing Office.

National Institutes of Health Toolbox for the Assessment of Neurological and Behavioral Function (NIH Toolbox). (2012). NIH Toolbox Emotion Battery: Domain-Specific Life Satisfaction Survey *(Supplemental Battery)* [measurement instrument]. NIH Toolbox. Available from http://www.healthmeasures.net/explore-measurement-systems/nih-toolbox.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: The National Academies Press. https://doi.org/10.17226/4962

Parker, J.G., and Asher, S.R. (1993). Friendship and Friendship Quality in Middle Childhood: Links with Peer Group Acceptance and Feelings of Loneliness and Social Dissatisfaction. *Developmental Psychology, 29*(4): 611-621.

Pianta, R.C., and Stuhlman, M.W. (2004). Teacher-Child Relationships and Children's Success in the First Years of School. *School Psychology Review, 33(*3): 444-458.

Pierce, K.M., Hamm, J.V., and Vandell, D.L. (1999). Experiences in After-School Programs and Children's Adjustment in First-Grade Classrooms. *Child Development, 70:* 756-767.

Princeton Survey Research Associates International. (2009). *PEW September Tracking Survey 2009.* Pew Internet & American Life Project.

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.

Raudenbush, S.W., and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition.* Newbury Park, CA: Sage.

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

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.

Rueda, M.R., Fan, J., McCandliss, B.D., Halparin, J.D., Gruber, D.B., Lercari, L.P., and Posner, M.I. (2004). Development of Attentional Networks in Childhood. *Neuropsychologia, 42*: 1029-1040.

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.

Singer, J.D., and Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Skinner, E.A., Kindermann, T.A., and Furrer, C.J. (2009). A Motivational Perspective on Engagement and Disaffection: Conceptualization and Assessment of Children's Behavioral and Emotional Participation in Academic Activities in the Classroom. *Educational and Psychological Measurement, 69*(3): 493-525.

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 August 20, 2014 from http://www.healthmeasures.net/images/nihtoolbox/Training-Admin-Scoring_Manuals/NIH_Toolbox_Scoring_and_Interpretation_Manual_9-27-12.pdf

Stocking, M., and Lord, F.M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement, 7*, 207–210.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A.G., Hagedorn, M.C., Daly, P., and Najarian, M. (2015a). *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). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Hagedorn, M.C., Leggitt, J., and Najarian, M. (2015b). *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). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L. and Najarian, M. (2015c). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Restricted Version* (NCES 2015-049). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L., and Najarian, M. (2017). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES 2017-285). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L. and Najarian, M. (2018a). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Third Grade Data File and Electronic Codebook, Public Version* (NCES 2018-034). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L. and Najarian, M. (2018b). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Fourth Grade Data File and Electronic Codebook, Public Version* (NCES 2018-032). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L. and Najarian, M. (2019). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten–Fifth Grade Data File and Electronic Codebook, Public Version* (NCES 2019-051). U.S. Department of Education. Washington, DC: National Center for Education Statistics

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. (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). Washington, DC: Author.

U.S. Department of Education, National Center for Education Statistics. (2019). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) Public-Use Kindergarten–Fifth Grade Data File and Electronic Codebook* (NCES 2019-050). Washington, DC: Author.

Vandell, D. (2000). *Peer Social Support, Bullying, and Victimization* (Form FLV05G3: *Kids in My Class at School)* [measurement instrument]. National Institute of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development: Phase III, 2000-2004. Available from http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21942.

Vandell, D. (2001). *Relationships With Peers: Part D (Teacher)* (Form FSV10G3: *Relationships with Peers: Teacher Version)* [measurement instrument]. National Institute of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development: Phase III, 2000-2004. Retrieved from http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21942.

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 Abilities.* Itasca, IL: Riverside Publishing.

Yamamoto, K., and Mazzeo, J. (1992). Item Response Theory Scale Linking in NAEP. *Journal of Educational Statistics*, *17*(2): 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.

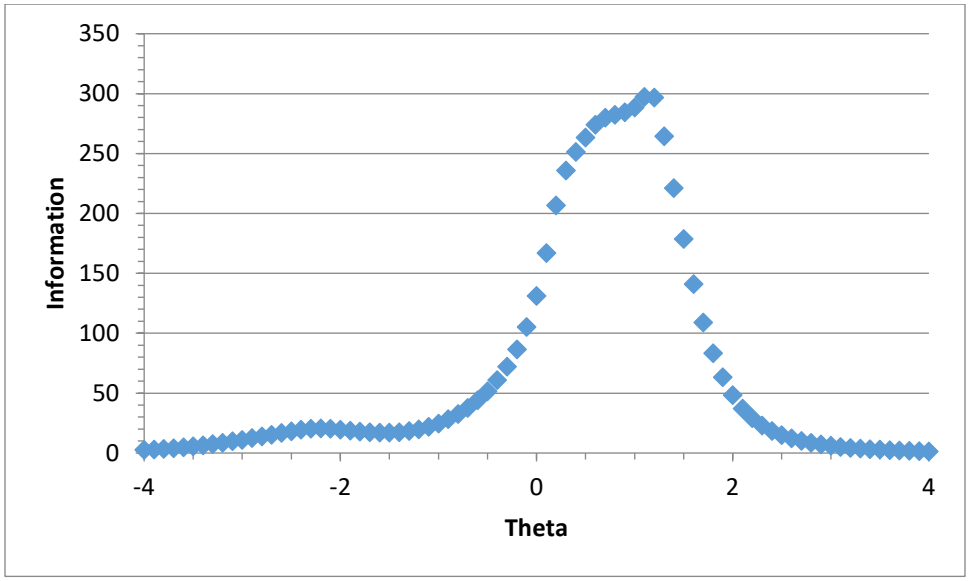
Zelazo, P.D., Anderson, J.E., Richler, J., Wallner-Allen, K., Beaumont, J.L., and Weintraub, S. (2013). NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. In P.D. Zelazo and P.J. Bauer (Eds.), *National Institutes of Health Toolbox Cognition Battery (NIH Toolbox CB): Validation for Children Between 3 and 15 Years. Monographs of the Society for Research in Child Development*.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement*. Princeton, NJ: Educational Testing Service.

APPENDIX A

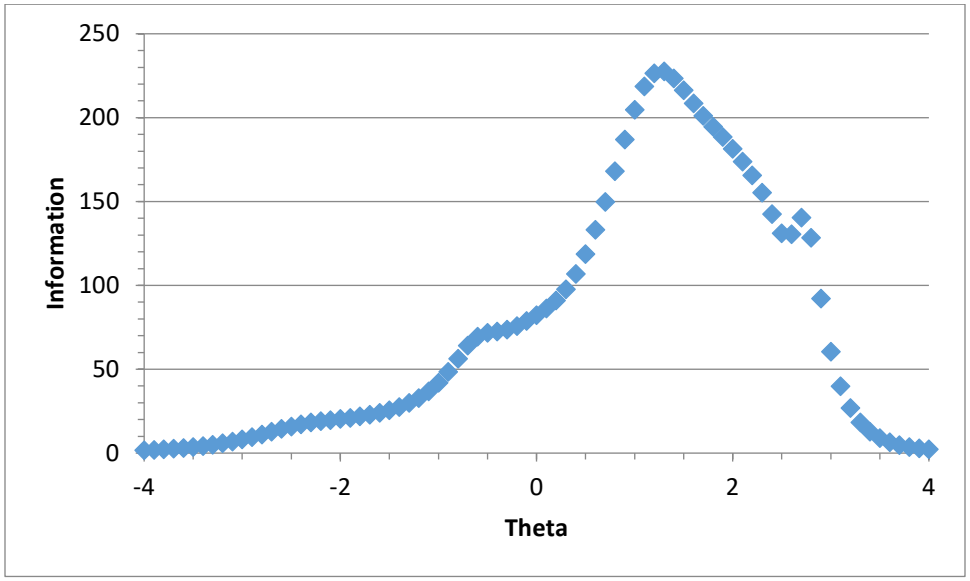
PLOTS OF ITEM RESPONSE THEORY (IRT) TEST INFORMATION FUNCTIONS FOR READING, MATHEMATICS, AND SCIENCE

Figure A-1. IRT test information function for reading for school years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016



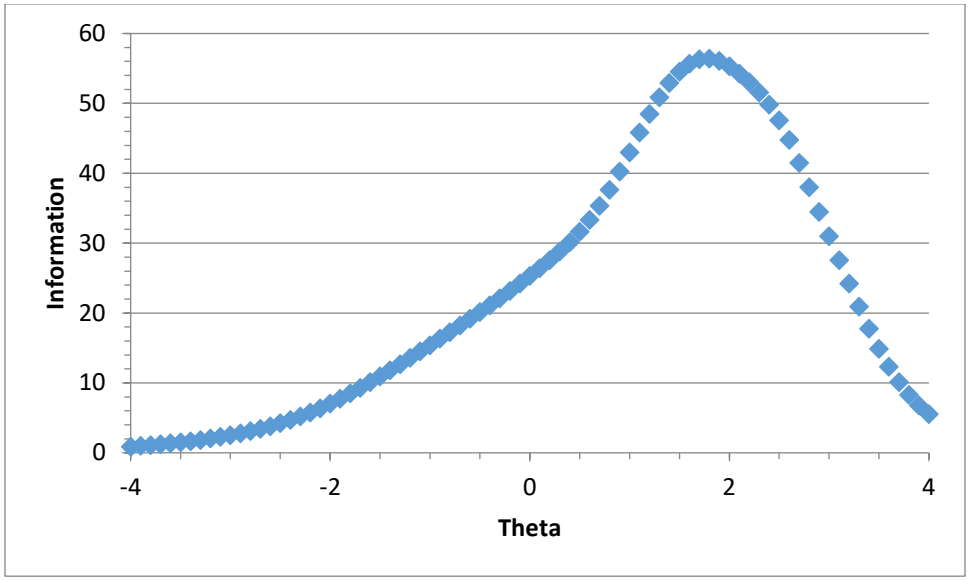
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), fall 2010, spring 2011, fall 2011, spring 2012, fall 2012, spring 2013, spring 2014, spring 2015, and spring 2016.

Figure A-2. IRT test information function for mathematics for school years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016



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, spring 2013, spring 2014, spring 2015, and spring 2016.

Figure A-3. IRT test information function for science for school years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016



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, spring 2013, spring 2014, spring 2015, and spring 2016.

APPENDIX B

ITEM RESPONSE THEORY (IRT) ESTIMATION USING PARSCALE

This appendix provides more detail on how the raw item responses are prepared for use in PARSCALE, how PARSCALE estimates the item response theory (IRT) model parameters, and what quality control checks are performed on the PARSCALE estimation output.

Preparing Data Files for PARSCALE

The first step in processing children's raw item responses was preparing scored-item files for use in the IRT calibration procedures. These files were first prepared separately for each round of data collection, fall kindergarten, spring kindergarten, fall first grade, spring first grade, fall second grade spring second grade, spring third grade, spring fourth grade, and spring fifth 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, kindergarten through second grade, and spring third, fourth, and fifth grades) 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. More formally, the general model

[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.

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:

$$\begin{aligned} L(\beta) &= \prod_g \prod_{j:g} \int_\theta P(x_{j:g} \mid \theta, \beta) f_g(\theta) d(\theta) \\ &\approx \prod_g \prod_{j:g} \Sigma_k P(x_{j:g} \mid \theta = X_k, \beta) A_g(X_k). \end{aligned} \quad (1)$$

In Equation 1, $L(\beta)$ is the marginalized likelihood of observing a given response matrix (students by items); $P(x_{j:g} \mid \theta, \beta)$ is the conditional probability of observing a response vector $\mathbf{x}_{j:g}$ 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, for the kindergarten through second-grade rounds, and singly for the spring third-, fourth-, and fifth-grade rounds. The present version of the multiple group PARSCALE saves the subpopulation means and standard deviations and the individual expected a posteriori (EAP) scores. The individual EAP scores, which are the means of the posterior distributions of theta,[2] were obtained using the Gaussian quadrature procedure. This procedure is virtually equivalent to conditioning (e.g., see Mislevy, Johnson, and Muraki 1992) on a set of dummy variables defining the ability subpopulation from which an observation comes. The one difference is that the group variances are not restricted to be equal as in the standard conditioning procedure.

Quality Control for PARSCALE Estimation

Statistics and graphs produced by the PARSCALE program and an IRT graphing program (PARPLOT) were used not only to verify the accuracy of the computations, but also to evaluate the reasonableness of the results. For each test item in the input scored data file, PARSCALE produced counts of the number of responses, number of omits, number right, number wrong, and percentage correct. These counts and percentages were checked, item by item, against the statistics generated from 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.

[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.

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).

APPENDIX C

ECLS-K:2011 KINDERGARTEN, FIRST-GRADE, SECOND-GRADE, THIRD-GRADE, FOURTH-GRADE, AND FIFTH-GRADE READING, MATHEMATICS, AND SCIENCE IRT ITEM PARAMETERS

Table C-1. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade reading IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]	d1[4]	d2[4]	d3[4]	d4[4]	d5[4]
MopMask	EBRS	EBRS	†	†	†	†	K-1-2	0.4707	-2.6584	0.0355	†	†	†	†	†
RunRope	EBRS	EBRS	†	†	†	†	K-1-2	0.5667	-1.6842	0.0161	†	†	†	†	†
ox_first	EBRS	EBRS	†	†	†	†	K-1-2	0.7509	-0.6096	0.0011	†	†	†	†	†
HE	EBRS	EBRS	†	†	†	†	K-1-2	2.3354	-0.5685	0.0014	†	†	†	†	†
goat_t	R2	R2	†	†	†	†	K-1-2	1.1626	-0.4016	0.0011	†	†	†	†	†
cat_c	R2	R2	†	†	†	†	K-1-2	1.7635	-0.4133	0.0011	†	†	†	†	†
plan_t	R2	R2	†	†	†	†	K-1-2	1.1269	-0.2469	0.0011	†	†	†	†	†
g_row	R2	R2	†	†	†	†	K-1-2	0.8994	0.3688	0.0011	†	†	†	†	†
red_snds	R2	R2	†	†	†	†	K-1-2	1.0885	-1.1605	0.0011	†	†	†	†	†
grab_snd	R2	R2	†	†	†	†	K-1-2	0.7096	0.4092	0.0011	†	†	†	†	†
RUNS	R2	R2	R2	†	†	†	K-1-2	2.5434	-0.2434	0.0011	†	†	†	†	†
WENT	R2	R2	R2	†	†	†	K-1-2	2.4209	-0.2083	0.0011	†	†	†	†	†
DOWN	R2	R2	R2	†	†	†	K-1-2	3.0491	-0.0458	0.0013	†	†	†	†	†
JEEP	R2	R2	R2	†	†	†	K-1-2	2.3109	0.0903	0.0012	†	†	†	†	†
boybike	R2	R2	R2	†	†	†	K-1-2	2.5254	-0.0106	0.1543	†	†	†	†	†
CANINBAG	R2	R2	R2	†	†	†	K-1-2	1.7325	0.0036	0.2255	†	†	†	†	†
kitnbed	R2	R2	R2	†	†	†	K-1-2	2.4246	0.0392	0.1601	†	†	†	†	†
girlread	R2	R2	R2	†	†	†	K-1-2	2.4556	0.2489	0.2076	†	†	†	†	†
GFCEREAL	L	†	†	†	†	†	K-1-2	1.6121	-2.9415	0.0011	†	†	†	†	†
which_s	L	†	†	†	†	†	K-1-2	1.1948	-2.3729	0.1446	†	†	†	†	†
BEGBIKE	L, M	L, M	†	†	†	†	K-1-2	1.1233	-2.4873	0.0011	†	†	†	†	†
Shelf_F	L, M, H	L, M, H	L, M, H	†	†	†	K-1-2	0.7996	-1.0867	0.0011	†	†	†	†	†
CANDLE	L	†	†	†	†	†	K-1-2	1.2071	-2.9292	0.3181	†	†	†	†	†
STATUE	L	†	†	†	†	†	K-1-2	1.5342	-2.3924	0.1272	†	†	†	†	†

See notes at end of table.

Table C-1. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade reading IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]	d1[4]	d2[4]	d3[4]	d4[4]	d5[4]
VEGETBLE	L	†	†	†	†	†	K-1-2	1.0754	-1.9399	0.1374	†	†	†	†	†
LIQUID	L, M	L, M	†	†	†	†	K-1-2	0.7892	-1.7069	0.0843	†	†	†	†	†
ROOT	L	†	†	†	†	†	K-1-2	1.1440	-1.6228	0.1814	†	†	†	†	†
TRUNK	M	M	†	†	†	†	K-1-2	0.3675	-1.1573	0.1166	†	†	†	†	†
PAIR	M	M	†	†	†	†	K-1-2	0.4716	0.1195	0.2028	†	†	†	†	†
Directng	M	M	†	†	†	†	K-1-2	0.4125	0.5786	0.2297	†	†	†	†	†
YOU	M	M	†	†	†	†	K-1-2	1.8926	-0.7434	0.0011	†	†	†	†	†
FROM	M	M	M	†	†	†	K-1-2	2.0518	-0.2419	0.0010	†	†	†	†	†
CATCH	M	M	†	†	†	†	K-1-2	2.3214	0.1463	0.0012	†	†	†	†	†
BACKPACK	H	H	H	†	†	†	K-1-2	3.4844	0.3437	0.1836	†	†	†	†	†
LISTEN	H	H	H	†	†	†	K-1-2	4.9864	0.4771	0.1799	†	†	†	†	†
RIDEBIKE	H	H	H	†	†	†	K-1-2	5.2207	0.5669	0.2196	†	†	†	†	†
SIZES	H	H	H	†	†	†	K-1-2	4.6867	0.6302	0.1783	†	†	†	†	†
before	H	H	H	†	†	†	K-1-2	5.7665	0.2140	0.0011	†	†	†	†	†
ALWAYS	H	H	H	†	†	†	K-1-2	5.2436	0.2757	0.0011	†	†	†	†	†
QUIET	H	H	H	†	†	†	K-1-2	3.7274	0.6263	0.0011	†	†	†	†	†
RAGE	H	H	H	†	†	†	K-1-2	3.1008	0.8938	0.0011	†	†	†	†	†
CL_LR	EBRS	EBRS	†	†	†	†	K-1-2	0.8874	-2.4144	0.0000	0.6264	-0.1986	-0.2292	-0.3406	0.1420
CL_LS	EBRS	EBRS	†	†	†	†	K-1-2	1.2420	-2.3682	0.0000	0.1799	-0.1799	†	†	†
CL_BS	R2	R2	†	†	†	†	K-1-2	0.4912	-1.5353	0.0000	0.3601	-0.3045	0.1167	-0.1724	†
CL_ES	R2	R2	†	†	†	†	K-1-2	0.6436	-0.7396	0.0000	0.4387	-0.1182	-0.3206	†	†
CL_BL	R2	R2	†	†	†	†	K-1-2	1.0839	-0.6226	0.0000	-0.0024	0.0024	†	†	†
CL_CS	R2	R2	†	†	†	†	K-1-2	0.9239	-0.2867	0.0000	0.1084	-0.1084	†	†	†
CL_RH	R2	R2	†	†	†	†	K-1-2	0.5204	0.2116	0.0000	-0.4340	0.4340	†	†	†
CL_WT	L	†	†	†	†	†	K-1-2	0.9748	-2.3392	0.0000	0.5731	-0.3287	-0.2444	†	†
CL_SO	L, M	L, M	†	†	†	†	K-1-2	0.3387	-2.4204	0.0000	-1.7048	1.7048	†	†	†
CL_SK	L, M	L, M	†	†	†	†	K-1-2	0.5658	-1.0270	0.0000	0.5786	-0.0024	-0.5762	†	†
CL_BR	H	H	H	†	†	†	K-1-2	2.5058	0.2168	0.0000	0.2746	0.1601	-0.0923	-0.3423	†
CL_DP	H	H	H	†	†	†	K-1-2	1.9648	0.8896	0.0000	0.3273	0.1047	-0.0840	-0.3480	†
CL_CM	H	H	H	†	†	†	K-1-2	1.9659	1.1268	0.0000	0.3133	0.1343	-0.1872	-0.2604	†
TOIL	†	R2	†	†	†	†	K-1-2	2.4926	0.8720	0.0011	†	†	†	†	†
UNUSUAL	†	R2	†	†	†	†	K-1-2	4.1597	0.9740	0.0012	†	†	†	†	†

See notes at end of table.

Table C-1. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade reading IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]	d1[4]	d2[4]	d3[4]	d4[4]	d5[4]
VICIOUS	†	R2	R	†	†	†	K-1-2	4.3229	1.2756	0.0013	†	†	†	†	†
KIMCAT	†	R2	R	†	†	†	K-1-2	3.3471	0.3253	0.4129	†	†	†	†	†
NEEDHOME	†	R2	R	†	†	†	K-1-2	1.9426	-0.1903	0.0197	†	†	†	†	†
LIKEDRY	†	R2	R	†	†	†	K-1-2	2.3579	0.5887	0.1354	†	†	†	†	†
TRAIN	†	H	M, H	†	†	†	K-1-2	4.3095	1.1811	0.2274	†	†	†	†	†
BOWMEAN	†	H	H	†	†	†	K-1-2	4.4192	1.1445	0.3471	†	†	†	†	†
TEARING	†	H	M, H	†	†	†	K-1-2	5.5552	1.2549	0.2996	†	†	†	†	†
CL_FR	†	R2	R	†	†	†	K-1-2	1.7959	0.5169	0.0000	0.5565	0.0523	-0.2037	-0.4051	†
CL_GG	†	H	M, H	†	†	†	K-1-2	2.5010	0.7960	0.0000	0.3241	0.0902	-0.1085	-0.3057	†
CL_SB	†	H	M, H	†	†	†	K-1-2	1.5141	1.0455	0.0000	0.5783	0.2704	-0.0411	-0.3250	-0.4826
CL_WJ	†	H	H	†	†	†	K-1-2	0.9290	1.3380	0.0000	0.3481	-0.0124	0.0130	-0.3487	†
FLATTIRE	†	†	R	†	†	†	K-1-2	2.8940	0.4278	0.1514	†	†	†	†	†
DOGHOUSE	†	†	R	†	†	†	K-1-2	2.2813	0.6296	0.0668	†	†	†	†	†
MARCHED	†	†	R	†	†	†	K-1-2	4.4778	0.8609	0.1900	†	†	†	†	†
MYSTERLY	†	†	R	†	†	†	K-1-2	4.4898	1.0540	0.0011	†	†	†	†	†
CRITICIS	†	†	R	†	†	†	K-1-2	3.4829	1.4598	0.0011	†	†	†	†	†
AMBITIOU	†	†	R	†	†	†	K-1-2	4.2090	1.4334	0.0011	†	†	†	†	†
CL_KA	†	†	R	†	†	†	K-1-2	1.7656	0.8518	0.0000	0.5073	0.2249	-0.011	-0.2543	-0.4671
CL_MF	†	†	†	R	†	M, H	3-4-5	1.6622	1.2623	0.0000	0.7529	0.4003	0.1850	-0.0142	-0.1902
CL_CH	†	†	†	R	R	R	3-4-5	1.2424	1.4804	0.0000	0.5814	0.1582	-0.0606	-0.6790	†
CL_RD	†	†	†	H	†	†	3-4-5	1.6761	1.2799	0.0000	0.4161	0.1268	-0.0776	-0.4653	†
CL_CC	†	†	†	H	H	M, H	3-4-5	1.9381	1.6044	0.0000	0.3446	0.1095	-0.0950	-0.3592	†
CL_MW	†	†	†	H	H	H	3-4-5	0.8694	2.5122	0.0000	0.9476	0.3104	-0.2362	-1.0217	†
Require	†	†	†	†	R	R	3-4-5	4.0103	0.8309	0.0000	†	†	†	†	†
Wages	†	†	†	†	R	R	3-4-5	3.1327	1.2659	0.0000	†	†	†	†	†
DOMESTIC	†	†	†	†	R	†	3-4-5	2.8928	1.1779	0.0000	†	†	†	†	†
Preferen	†	†	†	†	R	†	3-4-5	2.4023	1.3301	0.0000	†	†	†	†	†
Alignmnt	†	†	†	†	R	H	3-4-5	3.4697	1.6065	0.0000	†	†	†	†	†
Insuffic	†	†	†	†	R	R	3-4-5	2.4068	1.6524	0.0000	†	†	†	†	†
wtless	†	†	†	†	L	†	3-4-5	4.3036	0.7792	0.0000	†	†	†	†	†
Embolism	†	†	†	†	H	H	3-4-5	1.4381	2.3192	0.0000	†	†	†	†	†
CL_AE	†	†	†	†	M, H	†	3-4-5	1.7153	1.3867	0.0000	0.5202	0.3051	0.1320	-0.0455	-0.2389

See notes at end of table.

Table C-1. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade reading IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]	d1[4]	d2[4]	d3[4]	d4[4]	d5[4]
CL_BB	†	†	†	†	†	R	3-4-5	1.1612	0.8996	0.0000	0.6473	0.2883	0.0343	-0.2585	-0.7113
CL_KI	†	†	†	†	†	L	3-4-5	1.2575	0.9904	0.0000	0.8823	0.6420	0.2468	0.0266	-0.2916
JEEP	†	†	†	†	†	L	3-4-5	2.6207	0.2486	0.0000	†	†	†	†	†
CANINBAG	†	†	†	R	L	†	3-4-5	1.6636	0.1139	0.1696	†	†	†	†	†
girlread	†	†	†	R	L	L	3-4-5	1.6176	0.0060	0.0995	†	†	†	†	†
FROM	†	†	†	L	†	†	3-4-5	1.7225	-0.1819	0.0000	†	†	†	†	†
BACKPACK	†	†	†	L	†	†	3-4-5	3.9514	0.3648	0.0859	†	†	†	†	†
LISTEN	†	†	†	L	R	†	3-4-5	2.6379	0.3570	0.0354	†	†	†	†	†
RIDEBIKE	†	†	†	L	R	†	3-4-5	9.8257	1.1530	0.7500	†	†	†	†	†
SIZES	†	†	†	L	R	†	3-4-5	3.4224	0.5496	0.0958	†	†	†	†	†
ALWAYS	†	†	†	L	†	†	3-4-5	5.3774	0.3226	0.0000	†	†	†	†	†
QUIET	†	†	†	L	†	L	3-4-5	3.5656	0.6097	0.0000	†	†	†	†	†
CL_DP	†	†	†	L	L	†	3-4-5	2.0411	0.8574	0.0000	0.3986	0.0960	-0.1249	-0.3697	†
TRAIN	†	†	†	L, M	L, M	L, M	3-4-5	4.9507	1.1245	0.2985	†	†	†	†	†
BOWMEAN	†	†	†	M	M	M	3-4-5	4.6700	1.0579	0.1838	†	†	†	†	†
TEARING	†	†	†	L, M	L, M	L, M	3-4-5	5.4706	1.1609	0.2775	†	†	†	†	†
CL_GG	†	†	†	L, M	†	†	3-4-5	2.4063	0.7480	0.0000	0.3621	0.0678	-0.1175	-0.3123	†
CL_SB	†	†	†	L, M, H	L, M	L	3-4-5	2.0695	1.0945	0.0000	0.4687	0.1972	-0.0527	-0.2310	-0.3822
CL_WJ	†	†	†	M	M	M	3-4-5	0.9261	1.3635	0.0000	0.2818	-0.0665	0.0393	-0.2545	†
FLATTIRE	†	†	†	R	L	L	3-4-5	2.5740	0.3791	0.0539	†	†	†	†	†
DOGHOUSE	†	†	†	R	L	L	3-4-5	2.1445	0.5545	0.0326	†	†	†	†	†
AMBITIOU	†	†	†	†	R	†	3-4-5	3.5199	1.3505	0.0000	†	†	†	†	†
CL_KA	†	†	†	R	R	†	3-4-5	1.8000	0.7657	0.0000	0.5440	0.1799	-0.0335	-0.2178	-0.4726

†Not administered.
[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: 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. The item parameters are associated with what is noted in the Calibration column. Items that were administered in one or more of the kindergarten, first-grade, or second-grade rounds (K-1-2) and in one or more of the third-, fourth-, or fifth-grade rounds (3-4-5) will have two sets of item parameters listed in the table (see section 3.5.2).
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, spring 2013, spring 2014, spring 2015, and spring 2016.

Table C-2. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11; 2011–12; 2012–13; spring 2014; spring 2015; and spring 2016

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
square	R	†	†	†	†	†	K-1-2	0.2059	-6.4630	0.1486
NUMBER4	R	†	†	†	†	†	K-1-2	2.0509	-2.4607	0.0009
NUMBER 9	R	R	†	†	†	†	K-1-2	1.6846	-1.6708	0.0009
NUMBER23	R	R	R	†	†	†	K-1-2	1.7710	-0.8689	0.0009
STICKBAT	R	R	†	†	†	†	K-1-2	0.8145	-1.8010	0.1408
sm-lg-sm	R	R	†	†	†	†	K-1-2	1.3807	-0.2068	0.2692
count20	R	R	R	†	†	†	K-1-2	0.9442	-1.5229	0.0009
Bill_can	R	R	†	†	†	†	K-1-2	0.7444	-2.0240	0.0009
3RD LINE	R	R	R	†	†	†	K-1-2	1.3068	-0.8832	0.0009
_78910	R	R	R	†	†	†	K-1-2	1.6605	-1.1107	0.0009
51015_25	R	R	R	†	†	†	K-1-2	1.8892	-0.0623	0.0010
1_7	R	R	R	†	†	†	K-1-2	1.4709	-0.5734	0.0009
3_2 CARS	R	R	R	†	†	†	K-1-2	1.1198	-0.6905	0.0009
8_6crayn	R	†	†	†	†	†	K-1-2	1.1821	-0.5049	0.0009
2_5marbl	R	R	R	†	†	†	K-1-2	1.2519	-0.4138	0.0009
5-1orang	R	R	R	†	†	†	K-1-2	1.3244	-0.1873	0.1533
3_7penny	R	R	R	†	†	†	K-1-2	1.8251	-0.0713	0.0011
11animal	R	†	†	†	†	†	K-1-2	1.1980	0.0542	0.0009
2CRAYONS	L	†	†	†	†	†	K-1-2	0.6741	-4.0397	0.0009
count10	L	†	†	†	†	†	K-1-2	1.1003	-2.7448	0.0009
3BANANAS	L	†	†	†	†	†	K-1-2	0.6175	-2.6943	0.2037
6BANANAS	L	†	†	†	†	†	K-1-2	0.9644	-1.4219	0.0009
fingers4	L	†	†	†	†	†	K-1-2	1.9125	-2.3793	0.0009
penny7	L	†	†	†	†	†	K-1-2	1.4893	-1.3968	0.1431
morheart	L	†	†	†	†	†	K-1-2	0.4785	-2.0543	0.3320
fewcooky	L	†	†	†	†	†	K-1-2	0.4521	-1.2995	0.0009
more12	L	†	†	†	†	†	K-1-2	1.1379	-1.6261	0.0009
fewmore	L	†	†	†	†	†	K-1-2	0.6359	-0.7089	0.0009
_choc	L,M	L	†	†	†	†	K-1-2	1.0631	-1.9355	0.0009

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11; 2011–12; 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
_vanilla	L,M	L	†	†	†	†	K-1-2	1.0785	-2.2020	0.0009
STRAW	L	†	†	†	†	†	K-1-2	1.1121	-2.5541	0.0009
3-1PENCL	L,M	L	†	†	†	†	K-1-2	0.9185	-1.6846	0.0009
NUMBER7	L	†	†	†	†	†	K-1-2	1.8029	-2.1035	0.0009
number17	L,M,H	L,M	L	†	†	†	K-1-2	1.5227	-1.1270	0.0009
pntbrush	L,M	L	†	†	†	†	K-1-2	1.1383	-1.6637	0.2541
number18	M	L	†	†	†	†	K-1-2	1.2142	-1.0679	0.1974
morecats	M	L	†	†	†	†	K-1-2	0.7555	-1.0497	0.1296
apples_7	M	L	†	†	†	†	K-1-2	0.8258	-0.5815	0.1081
000X	M,H	L,M	L	†	†	†	K-1-2	1.0637	-0.4231	0.2379
trapzoid	M	L	†	†	†	†	K-1-2	0.7749	-0.9446	0.1200
HALFOVAL	M	L	†	†	†	†	K-1-2	0.8973	-1.0838	0.1904
crayons6	M	L	†	†	†	†	K-1-2	0.7520	-1.3245	0.0843
6_2 cars	M	L	†	†	†	†	K-1-2	1.2837	-0.4508	0.0009
2_2butrf	M	L	†	†	†	†	K-1-2	1.3305	-0.6625	0.1182
3-1books	M	L	†	†	†	†	K-1-2	0.8464	-0.9304	0.0510
bugs100	M	L	†	†	†	†	K-1-2	1.0978	-0.6399	0.1570
2_2	M,H	L,M	L	†	†	†	K-1-2	3.0887	-0.6801	0.0009
3_3	M	L	†	†	†	†	K-1-2	4.1515	-0.6075	0.0009
3_4	M,H	L,M	L	†	†	†	K-1-2	2.1243	-0.4092	0.0009
26_20	M,H	L,M	L	†	†	†	K-1-2	2.4930	0.5832	0.0009
12 BY 2S	M	L	†	†	†	†	K-1-2	1.7303	-0.0911	0.0009
smtolrg_	M	L	†	†	†	†	K-1-2	1.8304	-0.1633	0.0009
2_3stick	H	M	L	†	†	†	K-1-2	1.0030	-0.8132	0.0009
11_3	H	M	L	†	†	†	K-1-2	2.0090	-0.0615	0.0009
12_6	H	M	L	†	†	†	K-1-2	1.9679	0.1558	0.0009
extpatrn	H	M	L	†	†	†	K-1-2	0.8541	-0.7884	0.0891
pattby2s	H	M	L	†	†	†	K-1-2	2.3181	-0.1026	0.0009
pattrn55	H	M	L	†	†	†	K-1-2	2.6764	0.0134	0.0009

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11; 2011–12; 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
13_79	H	M,H	L,M	†	†	†	K-1-2	1.3508	0.3078	0.0009
_MORE	H	M	L	†	†	†	K-1-2	1.9057	0.6816	0.0009
2_5CIRCL	H	M	L	†	†	†	K-1-2	1.3844	-0.2966	0.0009
4_4-2	H	M	L	†	†	†	K-1-2	1.4688	0.6611	0.0009
17-4	H	M	L	†	†	†	K-1-2	2.0871	0.5997	0.0009
12-9	H	M	L	†	†	†	K-1-2	2.0426	0.5986	0.0009
COST_10	H	M,H	L,M	†	†	†	K-1-2	1.9516	0.6021	0.0009
15_5CARS	H	M,H	L,M	†	†	†	K-1-2	2.2758	0.7965	0.0009
8_2CANDY	H	M,H	L,M	†	†	†	K-1-2	1.9667	0.9121	0.0010
HOWMANY_	H	M,H	L,M	†	†	†	K-1-2	1.8499	1.0203	0.0009
12-_pen	H	M,H	L,M	†	†	†	K-1-2	1.8523	0.9591	0.0009
headsup	H	M,H	L,M	†	†	†	K-1-2	1.0181	1.1483	0.0009
24-14BKS	H	M,H	L,M	†	†	†	K-1-2	2.3662	1.0254	0.0009
8_4block	H	M,H	L,M	†	†	†	K-1-2	1.3577	1.0723	0.0009
nickels	H	M,H	L,M	†	†	†	K-1-2	1.8041	1.3005	0.0011
17CENTS	H	M,H	L,M	†	†	†	K-1-2	2.5268	1.1649	0.0009
2x5kids	†	R	†	†	†	†	K-1-2	2.2775	0.7939	0.0010
FRIES43	†	R	R	†	†	†	K-1-2	3.1401	1.5889	0.0010
CARDS579	†	R	R	†	†	†	K-1-2	2.7742	1.8927	0.0009
number60	†	H	M,H	†	†	†	K-1-2	3.3098	1.1131	0.0009
NEXT_R	†	H	M,H	†	†	†	K-1-2	2.7963	1.2751	0.0009
chart64	†	H	M,H	†	†	†	K-1-2	1.7892	1.1277	0.0009
box700	†	H	M,H	†	†	†	K-1-2	3.5622	1.2789	0.0009
allcans6	†	H	M	†	†	†	K-1-2	1.4335	0.7573	0.0737
weigh9_5	†	H	M,H	†	†	†	K-1-2	1.2238	1.3550	0.0009
pretzels	†	H	M	†	†	†	K-1-2	1.5314	1.1493	0.0009
samesize	†	H	M	†	†	†	K-1-2	1.2585	1.1973	0.0468
63plants	†	H	M,H	†	†	†	K-1-2	2.2168	1.5101	0.2318
fruit3_6	†	H	M,H	†	†	†	K-1-2	1.9507	1.6980	0.1143

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11; 2011–12; 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
CHARGE_Y	†	H	M,H	†	†	†	K-1-2	1.9758	1.7704	0.0009
MARIA_B	†	H	M,H	†	†	†	K-1-2	2.7852	1.6723	0.0009
cupcakes	†	H	M	†	†	†	K-1-2	1.8651	1.3912	0.0009
probblue	†	H	M,H	†	†	†	K-1-2	1.2644	1.6135	0.1417
8markers	†	H	M	†	†	†	K-1-2	1.7442	1.8287	0.0009
lrgfrac	†	H	M	†	†	†	K-1-2	1.2274	2.2089	0.0009
walkschl	†	H	M	†	†	†	K-1-2	2.4592	1.6460	0.1472
72window	†	H	M,H	†	†	†	K-1-2	1.7796	1.6921	0.0009
bar_pie	†	†	R	†	†	†	K-1-2	1.4409	0.1115	0.3194
8_17	†	†	R	†	†	†	K-1-2	1.3121	0.3807	0.0009
SIDES_R	†	†	R	†	†	†	K-1-2	1.4048	0.8292	0.1958
FEWEST_Y	†	†	R	†	†	†	K-1-2	1.4972	0.7280	0.0009
notebkpg	†	†	R	†	†	†	K-1-2	0.8414	0.6726	0.0895
recycle	†	†	R	†	†	†	K-1-2	1.8551	1.4729	0.0009
fracordr	†	†	R	†	†	†	K-1-2	0.7705	1.9022	0.0009
marbles	†	†	R	†	†	†	K-1-2	2.3992	2.1749	0.0010
PAPER	†	†	H	†	†	†	K-1-2	1.9458	1.0234	0.0009
9_5_2x2	†	†	H	†	†	†	K-1-2	2.0564	1.1444	0.2403
wheels	†	†	H	†	†	†	K-1-2	1.3865	1.2411	0.0009
change71	†	†	H	†	†	†	K-1-2	2.3372	1.5645	0.0009
GREW4	†	†	H	†	†	†	K-1-2	2.5326	1.8611	0.0009
MINUTE_Y	†	†	H	†	†	†	K-1-2	2.8631	1.8323	0.0009
MONEY_B	†	†	H	†	†	†	K-1-2	2.7150	1.8781	0.0009
mystery	†	†	H	†	†	†	K-1-2	2.2889	2.2631	0.0009
SAME_B	†	†	H	†	†	†	K-1-2	1.4321	2.4771	0.0009
colorred	†	†	†	R	†	†	3-4-5	1.3956	0.7672	0.0002
PAGES_R	†	†	†	R	†	†	3-4-5	1.7310	0.6226	0.3480
NEXT78	†	†	†	R	†	†	3-4-5	2.2311	1.0120	0.0002
TIME	†	†	†	R	R	R	3-4-5	1.6679	1.3310	0.0002

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11; 2011–12; 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
Flag	†	†	†	R	R	R	3-4-5	2.2019	1.3614	0.2414
24LEMONS	†	†	†	R	R	R	3-4-5	2.5313	1.6362	0.0003
LONGSTEP	†	†	†	R	†	†	3-4-5	0.8559	1.9730	0.0002
HwTIME	†	†	†	R	R	†	3-4-5	4.0731	1.9011	0.3782
number5	†	†	†	R	R	†	3-4-5	3.4962	2.4249	0.0764
Angle	†	†	†	R	R	R	3-4-5	1.4997	2.6090	0.1312
carpet	†	†	†	R	R	R	3-4-5	2.3812	2.5884	0.0003
radius	†	†	†	R	†	†	3-4-5	0.9350	4.6993	0.0002
SPOONS	†	†	†	H	M	M	3-4-5	2.0947	1.2844	0.0002
SCALE_	†	†	†	H	M	M	3-4-5	2.3785	1.6819	0.0002
teaspoon	†	†	†	H	M,H	†	3-4-5	3.3277	2.0741	0.0002
EDGES_B	†	†	†	H	M	M	3-4-5	1.3616	2.1339	0.0002
segment	†	†	†	H	M	M	3-4-5	2.1402	2.1221	0.1056
area	†	†	†	H	M,H	M,H	3-4-5	4.3018	2.2458	0.1311
sameangl	†	†	†	H	M,H	H	3-4-5	2.9616	2.5848	0.1864
MEASDIAM	†	†	†	H	H	H	3-4-5	2.8215	2.5643	0.0002
CircCirc	†	†	†	H	H	H	3-4-5	5.5878	2.7642	0.1030
FoldCube	†	†	†	H	†	H	3-4-5	2.6111	2.8332	0.1120
More_1	†	†	†	†	R	†	3-4-5	2.2662	1.2592	0.0002
LowDive	†	†	†	†	R	H	3-4-5	2.0539	2.2814	0.0902
num129	†	†	†	†	R	†	3-4-5	1.9922	1.9140	0.0002
12inch	†	†	†	†	R	H	3-4-5	3.0045	2.7690	0.0002
paraline	†	†	†	†	L	L	3-4-5	1.1628	0.3940	0.0002
number14	†	†	†	†	L	†	3-4-5	1.9320	0.8477	0.0002
bloodtyp	†	†	†	†	H	†	3-4-5	3.1193	2.0746	0.2148
sandwich	†	†	†	†	H	R	3-4-5	2.2722	1.9790	0.0002
shade	†	†	†	†	H	R	3-4-5	3.3065	2.3720	0.0003
tabletru	†	†	†	†	H	†	3-4-5	1.6277	2.5517	0.1780
range	†	†	†	†	H	H	3-4-5	2.8058	2.7571	0.2362

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11; 2011–12; 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
SamJuan	†	†	†	†	†	R	3-4-5	2.6591	1.3025	0.0792
pgsleft	†	†	†	†	†	R	3-4-5	3.0528	1.4410	0.1602
100more	†	†	†	†	†	R	3-4-5	2.2850	1.6907	0.0841
fourline	†	†	†	†	†	R	3-4-5	1.9067	2.1494	0.2781
num129	†	†	†	†	†	R	3-4-5	1.8627	1.9329	0.0002
xX5PLUS3	†	†	†	†	†	R	3-4-5	3.1014	2.8424	0.0866
vanillapc	†	†	†	†	†	M	3-4-5	2.6469	2.0797	0.0489
rhombus	†	†	†	†	†	M	3-4-5	1.7911	2.3597	0.2225
BUDGETFR	†	†	†	†	†	M,H	3-4-5	2.6662	1.9367	0.2603
compare	†	†	†	†	†	M	3-4-5	3.4447	2.2477	0.1651
7xplus4	†	†	†	†	†	H	3-4-5	2.4559	2.4055	0.2417
diff88	†	†	†	†	†	H	3-4-5	3.8708	2.4245	0.1225
twoequal	†	†	†	†	†	H	3-4-5	4.4278	2.7146	0.1106
radiusft	†	†	†	†	†	H	3-4-5	7.3639	2.7379	0.1302
51015_25	†	†	†	R	†	†	3-4-5	1.7994	-0.4112	0.0002
5-1orang	†	†	†	†	L	†	3-4-5	1.4764	0.1918	0.1746
26_20	†	†	†	†	R	†	3-4-5	1.5263	-0.0047	0.0002
12_6	†	†	†	†	R	†	3-4-5	1.1885	-0.0665	0.0002
pattrn55	†	†	†	R	L	L	3-4-5	2.1314	-0.0732	0.0002
13_79	†	†	†	L	†	L	3-4-5	1.1417	0.2923	0.0002
17-4	†	†	†	†	L	L	3-4-5	1.3097	0.4492	0.0002
12-9	†	†	†	†	†	L	3-4-5	1.7054	0.7542	0.0002
COST_10	†	†	†	L	L	†	3-4-5	2.4983	0.7459	0.0002
15_5CARS	†	†	†	L	†	†	3-4-5	2.8422	0.8918	0.0002
8_2CANDY	†	†	†	L	R	R	3-4-5	2.4108	0.9885	0.0002
HOWMANY_	†	†	†	L	†	†	3-4-5	1.9992	1.1016	0.0002
12-_pen	†	†	†	L	†	†	3-4-5	2.1288	1.1230	0.0002
headsup	†	†	†	L	†	†	3-4-5	1.3359	1.0528	0.0002
24-14BKS	†	†	†	L	†	†	3-4-5	2.9091	1.0286	0.0002

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11; 2011–12; 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
8_4block	†	†	†	L	†	†	3-4-5	1.9238	1.0539	0.0002
nickels	†	†	†	L	†	†	3-4-5	2.0919	1.3641	0.0002
17CENTS	†	†	†	L	†	†	3-4-5	2.9735	1.2932	0.0002
2x5kids	†	†	†	†	†	L	3-4-5	2.1813	0.7636	0.0002
number60	†	†	†	L,M	L,M	†	3-4-5	3.1583	1.1412	0.0002
NEXT_R	†	†	†	L,M	L,M	L,M	3-4-5	2.7452	1.2310	0.0002
chart64	†	†	†	L,M	†	†	3-4-5	1.7210	1.2963	0.0002
box700	†	†	†	L,M	L	†	3-4-5	3.2933	1.1672	0.0002
allcans6	†	†	†	L	L,M	L	3-4-5	1.1747	1.0393	0.0658
weigh9_5	†	†	†	L,M,H	L,M	L,M	3-4-5	1.3844	1.3490	0.0002
pretzels	†	†	†	L	†	†	3-4-5	1.9184	1.1621	0.0002
samesize	†	†	†	L	†	†	3-4-5	1.7082	1.3079	0.0540
63plants	†	†	†	L,M,H	M	M	3-4-5	1.5459	1.8047	0.1403
fruit3_6	†	†	†	L,M,H	L,M	L,M	3-4-5	1.8736	1.2622	0.2339
MARIA_B	†	†	†	L,M,H	L,M	L,M	3-4-5	2.5390	1.6617	0.0002
cupcakes	†	†	†	L	†	†	3-4-5	1.7169	1.1119	0.0002
CHARGE_Y	†	†	†	L,M	L	L,M,H	3-4-5	2.6317	1.5948	0.0002
probblue	†	†	†	L,M	L	L	3-4-5	0.9994	1.0733	0.1132
72window	†	†	†	M	H	†	3-4-5	2.0439	1.4377	0.0002
bar_pie	†	†	†	R	†	†	3-4-5	1.3187	-0.1839	0.1142
FEWEST_Y	†	†	†	R	R	R	3-4-5	1.9502	0.7595	0.0002
recycle	†	†	†	R	R	R	3-4-5	1.4439	1.6251	0.0002
marbles	†	†	†	†	R	R	3-4-5	2.9160	2.0021	0.0003
PAPER	†	†	†	M	†	†	3-4-5	2.2343	1.1299	0.0002
9_5_2x2	†	†	†	M	L,M,H	L	3-4-5	1.8871	0.9459	0.1394
wheels	†	†	†	M	L	L	3-4-5	2.2873	1.1449	0.0002
change71	†	†	†	M	L	L	3-4-5	2.5000	1.5550	0.0002
GREW4	†	†	†	M,H	L,M,H	L,M,H	3-4-5	2.4006	1.8949	0.0003
MINUTE_Y	†	†	†	M,H	L,M,H	L,M,H	3-4-5	2.4633	1.9548	0.0002

See notes at end of table.

Table C-2. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade mathematics IRT item parameters on the kindergarten scale: School years 2010–11; 2011–12; 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
MONEY_B	†	†	†	M,H	M,H	†	3-4-5	2.1986	2.0906	0.0002
mystery	†	†	†	M	H	R	3-4-5	2.7570	2.2423	0.0003
SAME_B	†	†	†	M	†	†	3-4-5	1.0829	2.4737	0.0002

†Not administered.
[1] Item Response Theory (IRT) discrimination parameter.
[2] Item Response Theory (IRT) difficulty parameter.
[3] Item Response Theory (IRT) guessing parameter.
NOTE: R = routing form, L = low second-stage form, M = middle second-stage form, H = high second-stage form. The item parameters are associated with what is noted in the Calibration column. Items that were administered in one or more of the kindergarten, first-grade, or second-grade rounds (K-1-2) and in one or more of the third-, fourth-, or fifth-grade rounds (3-4-5) will have two sets of item parameters listed in the table (see section 3.5.2).Additionally, one item (num129) appeared in both the fourth-grade and fifth-grade forms. However, preliminary analysis indicated that this item should not be treated as a common item across rounds, Therefore, it has two sets of item parameters listed in the table, one set for each grade in which it was included.
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, spring 2013, spring 2014, spring 2015, and spring 2016.

Table C-3. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade science IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
crabeach	R	†	†	†	†	†	K-1-2	1.4235	-1.5174	0.7500
nail	R	†	†	†	†	†	K-1-2	0.7268	-0.8109	0.5626
winter	R	†	†	†	†	†	K-1-2	0.6485	-2.0215	0.3594
roundobj	R	L, M	L	†	†	†	K-1-2	0.8058	-1.4827	0.0000
beehoney	R	R	R	†	†	†	K-1-2	1.2572	-1.5838	0.0000
spole	R	†	†	†	†	†	K-1-2	1.1370	-1.1762	0.0000
cloudsun	R	R	R	†	†	†	K-1-2	0.8054	-1.2167	0.0000
metalgrp	R	L, M, H	L, M	†	†	†	K-1-2	1.2603	-0.7694	0.1523
meltice	R	R	R	†	†	†	K-1-2	0.8876	-0.8700	0.1505
thermtr	R	L	†	†	†	†	K-1-2	0.9603	-1.3306	0.0000
froggrow	R	R	R	†	†	†	K-1-2	1.2070	-0.2439	0.0000
Snakemov	R	L	†	†	†	†	K-1-2	1.0215	-0.2616	0.3019
Eatplant	R	R	R	†	†	†	K-1-2	0.9248	0.0963	0.4359
Solidrck	R	R	R	†	†	†	K-1-2	1.0427	-0.2433	0.1369
glasses	R	R	R	†	†	†	K-1-2	0.6701	1.3033	0.1874
liqsolid	R	L	†	†	†	†	K-1-2	1.6243	0.4405	0.1620
hipitch	R	R	R	†	†	†	K-1-2	0.9586	0.9710	0.0306
poolwarm	R	R	†	†	†	†	K-1-2	0.9549	0.9235	0.1191
natresbx	R	M	L	†	†	†	K-1-2	1.0180	0.8907	0.2174
eatstom	R	R	†	†	†	†	K-1-2	0.9259	0.7179	0.2916
treeseas	†	R	R	†	†	†	K-1-2	0.6794	-0.9601	0.2828
leopard	†	R	†	†	†	†	K-1-2	1.1905	0.5276	0.0000
roots	†	R	†	†	†	†	K-1-2	0.9993	0.2435	0.0000
ROUGRT	†	R	†	†	†	†	K-1-2	1.5857	1.6421	0.0848
YLIVE	†	R	R	†	†	†	K-1-2	0.0262	34.2456	0.0056
garden	†	L	†	†	†	†	K-1-2	0.7269	-2.5640	0.0000
icemelts	†	L, M	L	†	†	†	K-1-2	0.8155	-1.9116	0.0000
strtfire	†	L, M	L	†	†	†	K-1-2	1.0141	-0.8052	0.1140
woodmetl	†	L, M	L	†	†	†	K-1-2	0.7350	-0.8112	0.1775

See notes at end of table.

Table C-3. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade science IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
airlungs	†	L, M, H	L, M	†	†	†	K-1-2	0.7738	-0.4565	0.0000
beakdrnk	†	L, M, H	L, M	†	†	†	K-1-2	1.0177	-0.4346	0.3284
sunnight	†	L, M	L	†	†	†	K-1-2	1.2374	0.5848	0.1061
ROUBRN	†	L, M, H	L, M	†	†	†	K-1-2	1.1619	0.5953	0.0000
Scale	†	M	L	†	†	†	K-1-2	0.8701	-0.3069	0.2828
owleat	†	M, H	L, M	†	†	†	K-1-2	0.7436	-0.2218	0.1320
liqgroup	†	M, H	L, M, H	†	†	†	K-1-2	1.3143	0.8873	0.1523
ROUSRF	†	M, H	L, M, H	†	†	†	K-1-2	1.4432	1.1540	0.4443
birdmigr	†	M	L	†	†	†	K-1-2	1.9981	1.1285	0.1424
dogdata	†	H	M	†	†	†	K-1-2	0.4372	0.1788	0.2585
YINSCT	†	H	M, H	L, M	†	†	K-1-2	1.3381	1.0304	0.1952
brightst	†	H	M, H	†	†	†	K-1-2	0.9341	1.7250	0.2828
shadow	†	H	M, H	†	†	†	K-1-2	1.0370	1.3073	0.0000
absorbs	†	H	M	†	†	†	K-1-2	1.3684	1.5461	0.0436
soundfst	†	H	M, H	†	†	†	K-1-2	1.8898	1.9295	0.1342
hibernat	†	H	M, H	†	†	†	K-1-2	1.1564	1.5059	0.2223
mammals	†	H	M, H	†	†	†	K-1-2	1.0534	2.5178	0.0000
allfoods	†	†	R	†	†	†	K-1-2	0.6875	-2.7689	0.0000
hshoemag	†	†	R		L	L	K-1-2	1.1010	-0.2971	0.0000
craymelt	†	†	R	†	†	†	K-1-2	0.4764	0.3882	0.0000
veggrow	†	†	R	†	†	†	K-1-2	1.6540	1.1906	0.3282
birdseed	†	†	R	†	†	†	K-1-2	0.9991	0.5203	0.2777
senses5	†	†	R	†	†	†	K-1-2	1.0797	1.4259	0.0000
solarcar	†	†	R	†	†	†	K-1-2	1.8063	1.2338	0.1398
leafsoil	†	†	R	†	†	†	K-1-2	1.0161	2.5295	0.0000
BPLLUT	†	†	R	†	†	†	K-1-2	1.0938	2.3755	0.1527
appltble	†	†	H	†	†	†	K-1-2	0.9504	0.7010	0.2120
rndsmth	†	†	H	†	†	†	K-1-2	0.9538	0.9390	0.2260
droplets	†	†	H	†	†	†	K-1-2	0.6898	1.7543	0.2324

See notes at end of table.

Table C-3. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade science IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
whyjar	†	†	H	†	†	†	K-1-2	0.6856	1.2666	0.0000
whichjar	†	†	H	†	†	†	K-1-2	1.2755	2.1706	0.2979
BPLNT3	†	†	H	†	†	†	K-1-2	1.5324	2.2862	0.1255
BSOIL	†	†	H	†	†	†	K-1-2	1.8691	2.5740	0.1600
oilwater	†	†	H	†	†	†	K-1-2	1.1437	2.3548	0.0000
ROCCUR	†	†	†	R	R	R	3-4-5	1.4223	0.1470	0.2886
S3110	†	†	†	R	R	R	3-4-5	1.9636	1.4597	0.2869
S5127	†	†	†	R	†	†	3-4-5	1.5441	1.1594	0.1224
S20019	†	†	†	R	R	R	3-4-5	1.8124	1.7736	0.3056
S6104	†	†	†	R	R	M, H	3-4-5	1.4106	2.3521	0.1322
Predator	†	†	†	R	†	†	3-4-5	2.7876	2.4993	0.1661
ROUJUN	†	†	†	L	†	L	3-4-5	1.3063	0.3619	0.0000
RTHING	†	†	†	L	R	R	3-4-5	1.0717	0.8772	0.2791
RANIML	†	†	†	M	†	†	3-4-5	1.4674	0.8418	0.3116
RSUNIS	†	†	†	M	L, M	L, M	3-4-5	0.7531	0.1944	0.0938
RPWDER	†	†	†	M	L, M	L	3-4-5	0.8240	0.9238	0.1853
S6130	†	†	†	M, H	†	†	3-4-5	1.4514	1.5097	0.1679
BURIED	†	†	†	M, H	R	R	3-4-5	1.1127	1.9586	0.2297
S20045	†	†	†	M, H	†	R	3-4-5	1.6687	2.2120	0.3052
S5126	†	†	†	M	†	†	3-4-5	1.5712	2.1356	0.2832
S20040	†	†	†	M	†	†	3-4-5	1.6068	2.5761	0.1136
ROUMCE	†	†	†	H	†	†	3-4-5	1.8831	1.4965	0.2061
S20030	†	†	†	H	M, H	M, H	3-4-5	1.8744	1.9222	0.2374
S6132	†	†	†	H	R	R	3-4-5	0.8861	2.3674	0.1627
S20049	†	†	†	H	†	R	3-4-5	1.2146	2.3698	0.2712
S3112	†	†	†	H	†	†	3-4-5	1.2040	2.2093	0.1323
S5122	†	†	†	H	H	H	3-4-5	1.3921	2.8888	0.3377
S20005	†	†	†	H	†	†	3-4-5	0.8027	3.8615	0.0000
airpoll	†	†	†		R	R	3-4-5	2.2146	1.5685	0.3906

See notes at end of table.

Table C-3. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade science IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
bugprefe	†	†	†	†	R	R	3-4-5	1.5815	2.4101	0.0000
buttrfly	†	†	†	†	L	L	3-4-5	1.2403	-0.1346	0.1333
prey	†	†	†	†	L	L	3-4-5	0.5865	-0.2494	0.0000
Shadow1	†	†	†	†	L, M, H	L, M	3-4-5	1.3953	1.6941	0.2849
Ohio	†	†	†	†	L, M, H	L, M, H	3-4-5	1.3783	1.8786	0.2568
YMOON	†	†	†	†	M	L, M	3-4-5	1.2037	1.2464	0.3740
RORGAN	†	†	†	†	M, H	M, H	3-4-5	1.0881	2.1536	0.1640
BPLNT	†	†	†	†	M, H	M, H	3-4-5	1.9208	2.0735	0.0658
CO2	†	†	†	†	M, H	M, H	3-4-5	1.5348	2.2579	0.0000
lens	†	†	†	†	H	H	3-4-5	2.1903	3.1670	0.0866
atom	†	†	†	†	H	H	3-4-5	2.2300	2.9685	0.0000
moonlook	†	†	†	†	H	H	3-4-5	1.6641	3.2662	0.2338
BSLIDE	†	†	†	†	†	R	3-4-5	1.2545	2.0226	0.1564
roundobj	†	†	†	†	R	†	3-4-5	0.7583	-1.5235	0.0000
metalgrp	†	†	†	L	L	†	3-4-5	1.0584	-1.0565	0.1658
froggrow	†	†	†	R	R	†	3-4-5	1.1510	0.0192	0.0000
Solidrck	†	†	†	R	†	†	3-4-5	0.9284	-0.2392	0.2510
leopard	†	†	†	†	L	†	3-4-5	1.1857	0.3340	0.0000
YLIVE	†	†	†	R	†	M, H	3-4-5	2.1919	1.7242	0.1258
woodmetl	†	†	†	†	L	L	3-4-5	0.8570	-0.6682	0.1824
airlungs	†	†	†	L	†	†	3-4-5	0.9417	-0.3524	0.0000
beakdrnk	†	†	†	L	L	L	3-4-5	0.7151	-0.5955	0.2082
sunnight	†	†	†	R	R	R	3-4-5	0.9890	0.6793	0.0881
ROUBRN	†	†	†	L	†	†	3-4-5	1.2147	0.7135	0.0000
Scale	†	†	†	†	R	†	3-4-5	0.4275	-1.1582	0.2118
owleat	†	†	†	L	†	†	3-4-5	0.8343	-0.1405	0.1608
liqgroup	†	†	†	L, M, H	L, M	L	3-4-5	1.4175	0.8182	0.1817
ROUSRF	†	†	†	L, M, H	†	†	3-4-5	1.9138	1.2018	0.4724
YINSCT	†	†	†	L, M	†	†	3-4-5	1.4646	1.2655	0.2445

See notes at end of table.

Table C-3. ECLS-K:2011 kindergarten, first-grade, second-grade, third-grade, fourth-grade, and fifth-grade science IRT item parameters on the kindergarten scale: School years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016—Continued

Item	Kinder-garten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Cali-bration	a[1]	b[2]	c[3]
brightst	†	†	†	L, M, H	M	M, H	3-4-5	0.9750	1.6166	0.2192
shadow	†	†	†	L, M	†	†	3-4-5	1.0175	1.4029	0.0000
absorbs	†	†	†	L	†	†	3-4-5	1.8453	1.3807	0.0469
mammals	†	†	†	H	H	M, H	3-4-5	1.0783	2.8290	0.0000
hshoemag	†	†	†	†	L	L	3-4-5	1.0720	-0.5239	0.0000
veggrow	†	†	†	L, M	L, M, H	L, M	3-4-5	1.5055	1.0793	0.4161
birdseed	†	†	†	R	†	†	3-4-5	0.8135	0.2003	0.1891
senses5	†	†	†	R	R	R	3-4-5	0.9853	1.5229	0.0000
solarcar	†	†	†	R	†	†	3-4-5	2.0284	1.2063	0.1857
leafsoil	†	†	†	R	R	R	3-4-5	1.4079	2.2105	0.0000
whichjar	†	†	†	M, H	†	†	3-4-5	0.7422	1.5874	0.1675
BPLNT3	†	†	†	†	†	R	3-4-5	1.7294	2.5867	0.0536
BSOIL	†	†	†	H	M, H	†	3-4-5	2.0187	2.0698	0.1201
oilwater	†	†	†	H	†	†	3-4-5	1.4172	2.2873	0.0000

† Not administered.
[1] Item Response Theory (IRT) discrimination parameter.
[2] Item Response Theory (IRT) difficulty parameter.
[3] Item Response Theory (IRT) guessing parameter.
NOTE: L = low second-stage form, M = middle second-stage form, H = high second-stage form. The item parameters are associated with what is noted in the Calibration column. Items that were administered in one or more of the kindergarten, first-grade, or second-grade rounds (K-1-2) and in one or more of the third-, fourth-, or fifth-grade rounds (3-4-5) will have two sets of item parameters listed in the table (see section 3.5.2).
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, spring 2013, spring 2014, spring 2015, and spring 2016.

APPENDIX D

PLOTS OF ITEM RESPONSE THEORY (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, 2012–13; spring 2014; spring 2015; and spring 2016

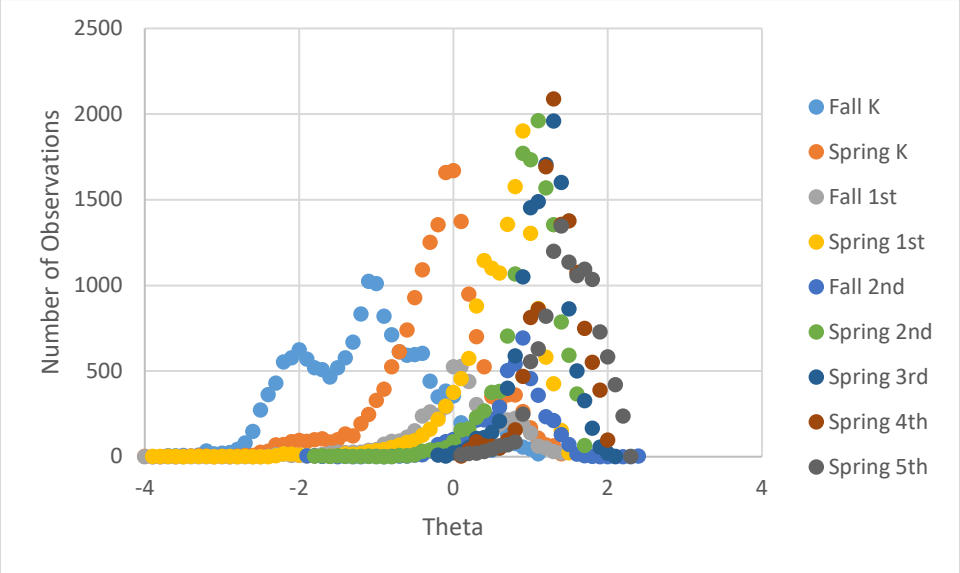


Figure D-2. IRT scale score distributions in reading for school years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

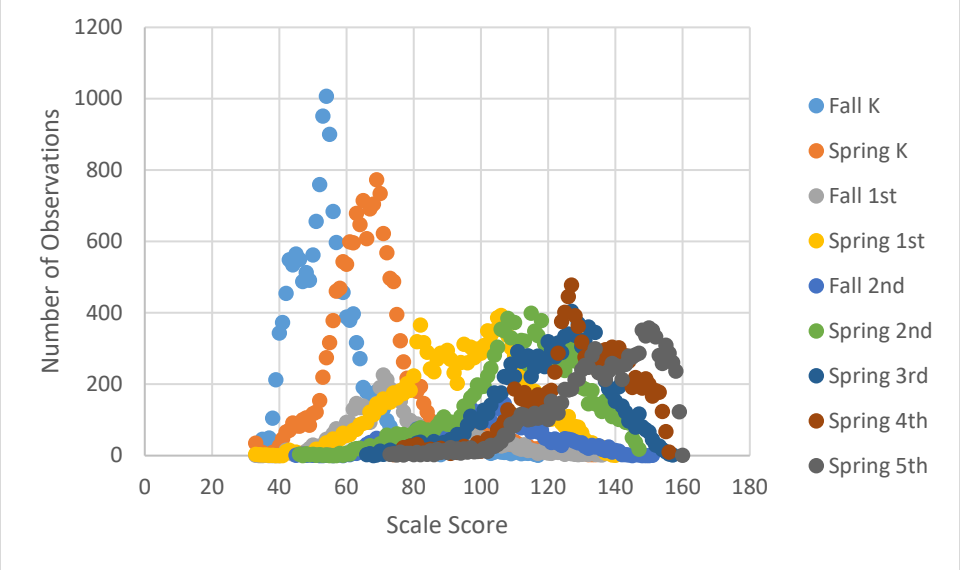


Figure D-3. IRT theta score distributions in mathematics for school years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

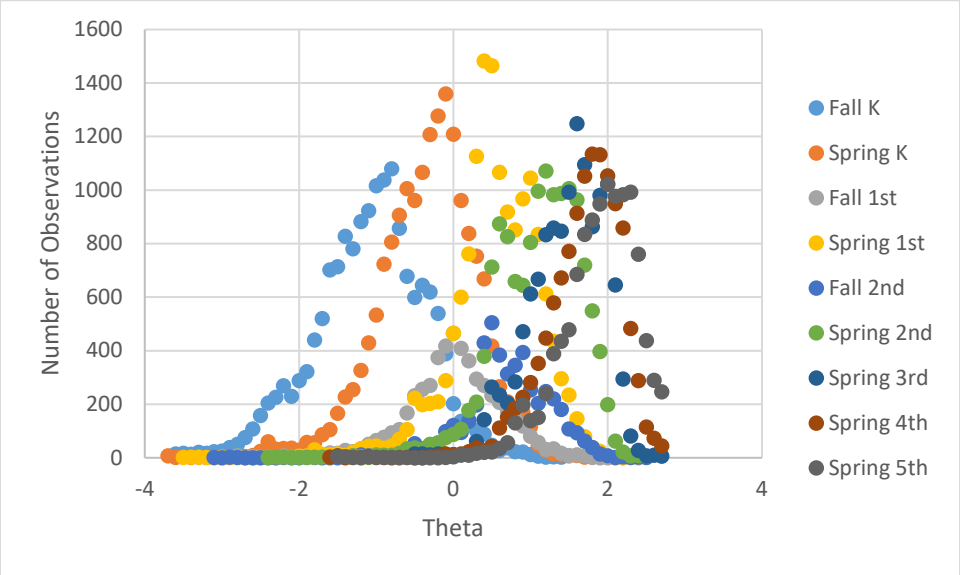


Figure D-4. IRT scale score distributions in mathematics for school years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

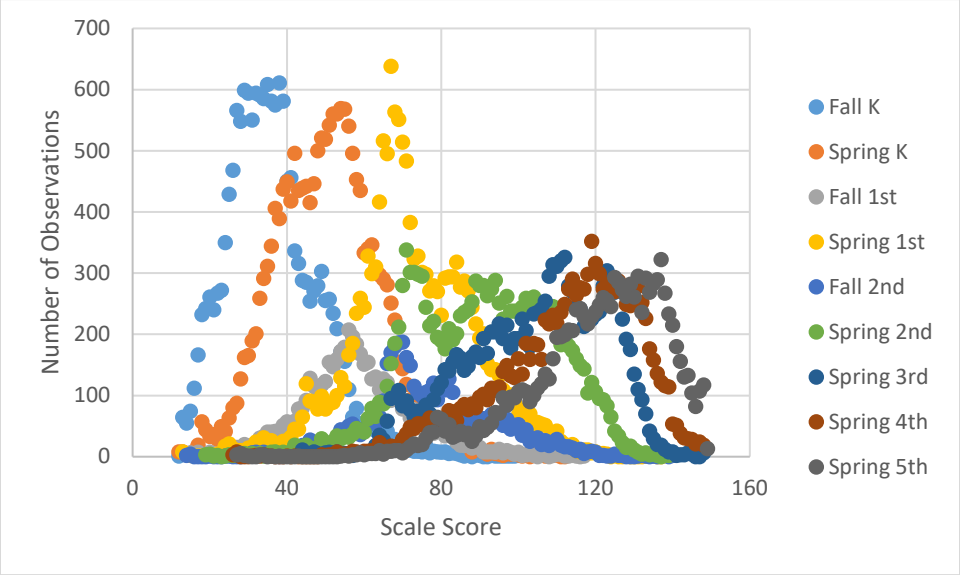


Figure D-5. IRT theta score distributions in science for school years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016

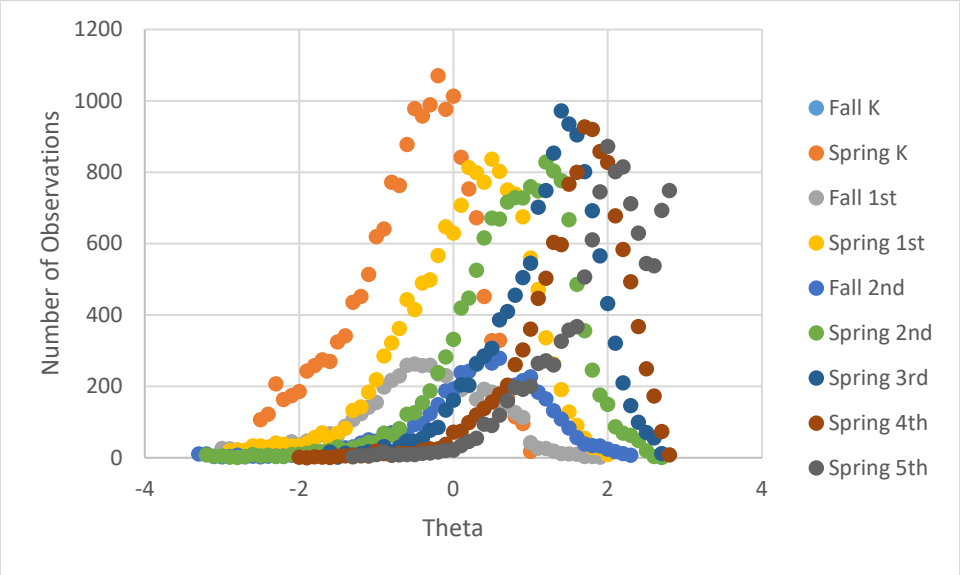


Figure D-6. IRT scale score distributions in science for school years 2010–11, 2011–12, 2012–13; spring 2014; spring 2015; and spring 2016