

Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K)

Psychometric Report for the Eighth Grade



Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K)

Psychometric Report for the Eighth Grade

September 2009

Michelle Najarian Judith M. Pollack Educational Testing Service

Alberto G. Sorongon Westat

Elvira Germino Hausken Project Officer National Center for Education Statistics U.S. Department of Education Arne Duncan Secretary

Institute of Education Sciences John Q. Easton *Director*

National Center for Education Statistics

Stuart Kerachsky Acting Commissioner

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

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

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

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

September 2009

The NCES World Wide Web Home Page address is <u>http://nces.ed.gov</u>. The NCES World Wide Web Electronic Catalog is <u>http://nces.ed.gov/pubsearch</u>.

This report is available only on the Web at http://nces.ed.gov/pubsearch.

Suggested Citation

Najarian, M. Pollack, J.M., and Sorongon, A.G. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Eighth Grade* (NCES 2009–002). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Content Contact Chris Chapman (202) 502-7414 chris.chapman@ed.gov

TABLE OF CONTENTS

Chapter				<u>Pag</u>
1	INTR	ODUCTI	ON	1-
2	DESI	GN AND	DEVELOPMENT OF THE ASSESSMENT	
	INST	RUMEN	TS	2-
	2.1	Direct	Cognitive Assessment	2-
		2.1.1	Administration of Adaptive Tests	2-
		2.1.2	The ECLS-K Frameworks	2-
			2.1.2.1 Reading Test Specifications	2-
			2.1.2.2 Mathematics Test Specifications	2-1
			2.1.2.3 Science Test Specifications	2-1
		2.1.3	Field Testing of Direct Cognitive Items	2-17
			2.1.3.1 Field Test Design	2-1
			2.1.3.2 Field Test Results and Conclusions	2-2
		2.1.4	Eighth-Grade Test Forms	2-2-
			2.1.4.1 Item Quality and Reliability	2-2
			2.1.4.2 Item Difficulty	2-2
			2.1.4.3 Floor and Ceiling Effects	2-2
			2.1.4.4 Longitudinal Score Scale	2-2
			2.1.4.5 Curriculum Relevance	2-2
			2.1.4.6 Framework Specifications	2-2
			2.1.4.7 Practical Issues	2-3
	2.2	Socioe	motional Development Measures	2-3
3	ANA	LYSIS M	ETHODOLOGY	3-
	3.1	Quality	y Control Procedures	3-
	3.2	Overvi	ew: The Three-Parameter Model	3-
		3.2.1	Overview of Item Response Theory	3-
		3.2.2	Item Response Theory Estimation Using PARSCALE	3-1
		3.2.3	Standard Errors of Measurement Using the Information Function	3-1
	33	Differe	ential Item Functioning	3_1
	5.5	Dinele		5-1

<u>Chapter</u>				Page
4	PSYC COGN	HOMETR NITIVE BA	LIC CHARACTERISTICS OF THE ECLS-K DIRECT ATTERY	4-1
	4.1	Types of	f Scores	4-1
		4.1.1 4.1.2	Number-Right Scores Item Response Theory Scale Scores; Standardized	4-2
			Scores (T-Scores)	4-3
		4.1.3	Item Cluster Scores	4-4
		4.1.4	Proficiency Levels	4-5
			4.1.4.1 Highest Proficiency Level Mastered	4-6
			4.1.4.2 Proficiency Probability Scores	4-8
	4.2	Reading	Assessment	4-9
		4.2.1 4.2.2	Samples and Operating Characteristics Scores Unique to the Reading Assessment:	4-9
			Proficiency Levels	4-11
		4.2.3	Reliabilities	4-12
		4.2.4	Score Statistics	4-15
		4.2.5	Standard Errors	4-16
		4.2.6	Differential Item Functioning	4-17
	4.3	Mathema	atics Assessment	4-18
		4.3.1 4.3.2	Samples and Operating Characteristics Scores Unique to the Mathematics Assessment:	4-18
			Proficiency Levels	4-21
		4.3.3	Reliabilities	4-22
		4.3.4	Score Statistics	4-24
		4.3.5	Standard Errors	4-25
		4.3.6	Differential Item Functioning	4-25
	4.4	Science .	Assessment	4-26
		4.4.1	Samples and Operating Characteristics	4-26
		4.4.2	Reliabilities	4-27
		4.4.3	Score Statistics	4-29
		4.4.4	Standard Errors	4-29
		4.4.5	Differential Item Functioning	4-30

<u>Chapter</u>				Page
5	DIREC MEAS	CT COGN UREMEN	ITIVE ASSESSMENTS: LONGITUDINAL NT	5-1
	5.1	Develop	ment of the K-1-3-5-8 Longitudinal Scale	5-1
		5.1.1 5.1.2 5.1.3	Second-Grade Bridge Study Evaluating Common Items IRT Calibration and Scoring	5-2 5-2 5-12
	5.2	Evaluati	ng the K-1-3-5-8 Longitudinal Scale	5-15
		5.2.1 5.2.2	Do the Tests Measure the Right Content? Is the Difficulty of the Tests Suitable for Children's	5-16
		5 2 3	Ability Levels?	5-17
		5.2.4	for Longitudinal Measurement? Relationship of the Cognitive Test Scores to Scores in	5-18
		525	Different Round and Different Subjects, and to Teacher Ratings and Child Self-Ratings Comparison of FCL S-K Results with Findings from	5-20
		5.2.5	Other Studies	5-22
	5.3	Applicat	ions	5-25
		5.3.1	Choosing Appropriate Scores for Analysis	5-26
			5.3.1.1 Item Response Theory-Based Scores5.3.1.2 Scores Based on Number Right for Subsets	5-26
			of Items (Non-IRT Based Scores)	5-27 5-28
		5.3.2	Notes on Measuring Gains	5-29
6	PSYCI SOCIO	HOMETR DEMOTIC	IC CHARACTERISTICS OF THE NAL MEASURES	6-1
	6.1	Self-Des	cription Questionnaire	6-1
		6.1.1 6.1.2 6.1.3	Reliability Analysis Factor Analysis Mean Scores	6-2 6-3 6-3

<u>Chapter</u>				Page
	6.2	Self-Con	cept and Locus of Control Scale Scores	6-7
		6.2.1 6.2.2	Reliability Analysis Factor Analysis	6-8 6-8
7	PSYC MEA	UNDETRI SURES	IC CHARACTERISTICS OF THE INDIRECT	7-1
	7.1	Teacher	Measures	7-1
		7.1.1	Indirect Cognitive Assessment Using the Academic Rating Scale (ARS)	7-2
			7.1.1.1 Floor and Ceiling	7-5
	7.2	Discrimit Indirect N	nant and Convergent Validity of the Direct and Measures	7-8
	REFE	RENCES		R-1
			List of Appendixes	
Appendix				
А	SCOF Roli	RE STATIS	TICS FOR DIRECT COGNITIVE MEASURES BY TA COLLECTION AND SELECTED	

	ROUND OF DATA COLLECTION AND SELECTED CHARACTERISTICS	A-1
В	ECLS-K ITEM PARAMETERS BY ROUND	B-1
С	ECLS-K ESTIMATED PROPORTION CORRECT BY ROUND	C-1
D	ECLS-K DIFFERENCE BETWEEN ACTUAL AND ESTIMATED PERCENT CORRECT BY ROUND	D-1

List of Tables

<u>Table</u>		
2-1	Reading longitudinal test specifications for kindergarten through eighth grade: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07.	
2-2	Mathematics longitudinal test specifications for kindergarten through eighth grade: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07.	
2-3	Science longitudinal test specifications, in percent of test items, for third grade (spring 2002), fifth grade (spring 2004), and eighth grade (spring 2007).	
2-4	Number of observations from the ECLS-K field test pool in field test forms, by section: Spring 2006 field test	
2-5	Reading eighth-grade framework targets and percent of assessment items: School year 2006-07	
2-6	Mathematics eighth-grade framework targets and percent of assessment items: School year 2006–07	
2-7	Science eighth-grade framework targets and percent of assessment items: School year 2006–07	
2-8	Number of items in eighth-grade test forms and routing test cut scores, by domain: School year 2006–07	
4-1	Reading assessment: Samples and operating characteristics: Rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006-07	
4-2	Reading assessment reliabilities, rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	
4-3	Reading assessment scale score means and standard deviations, rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	
4-4	Reading assessment mean theta score and mean standard error, rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	

<u>Table</u>		Page
4-5	Reading assessment: Differential item functioning, eighth grade: School year 2006–07	4-17
4-6	Mathematics assessment: samples and operating characteristics, rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	4-20
4-7	Mathematics assessment reliabilities, rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	4-22
4-8	Mathematics assessment scale score means and standard deviations, rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	4-24
4-9	Mathematics assessment mean theta score and mean standard error, rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	4-25
4-10	Mathematics assessment: Differential item functioning, eighth grade: School year 2006–07	4-26
4-11	Science assessment: Samples and operating characteristics, rounds 5 through 7: School years 2001–02, 2003–04, and 2006–07	4-27
4-12	Science assessment reliabilities, rounds 5 through 7: School years 2001–02, 2003–04, and 2006–07	4-28
4-13	Science scale score mean and standard deviation, rounds 5 through 7: School years 2001–02, 2003–04, and 2006–07	4-29
4-14	Science mean theta score and mean standard error, rounds 5 through 7: School years 2001–02, 2003–04, and 2006–07	4-30
4-15	Science assessment: Differential item functioning, eighth grade: School year 2006–07	4-30
5-1	Counts of common items, separate items, and total items in item pools: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	5-3
5-2	Reading assessment, actual minus predicted proportion correct: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	5-5

<u>Table</u>		Page
5-3	Mathematics assessment, actual minus predicted proportion correct: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	5-8
5-4	Science assessment, actual minus predicted proportion correct: School years 2001–02, 2003–04, and 2006–07	5-10
5-5	IRT theta (ability) means and standard deviations by subpopulation, seven data collection rounds plus bridge sample: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	5-13
5-6	IRT parameters for reading and mathematics proficiency levels, based on items from kindergarten, first-grade, third-grade, fifth-grade, and eighth-grade assessments: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	5-15
5-7	Correlations of IRT theta score across rounds, by subject: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	5-21
5-8	Correlations of IRT theta score across subjects, by round: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	5-22
5-9	Subgroup gaps in standard deviation units, NAEP and ECLS-K: School year 2006-07	5-24
6-1	Reliability estimates for scores of the self-description questionnaire scales, spring-eighth grade: School year 2006–07	6-2
6-2	Eigenvalues and proportion of variance accounted for by the three factors extracted in principal components factor analysis with self-description questionnaire data: School year 2006–07	6-3
6-3	Self-description questionnaire weighted means and standard deviations, spring-eighth grade: School year 2006–07	6-3
6-4	Score breakdown, self-description questionnaire, Perceived Interest/ Competence in Reading, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006–07	6-4
6-5	Score breakdown, self-description questionnaire, Perceived Interest/ Competence in Mathematics, by eighth-graders, sixth- and seventh- graders, and population subgroup: School year 2006–07	6-5

<u>Table</u>		Page
6-6	Score breakdown, self-description questionnaire, Internalizing Problems, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006–07	6-0
6-7	Self-Concept and the Locus of Control scale reliabilities (alpha coefficient): School year 2006–07	6-8
6-8	Eigenvalues and proportion of variance accounted for by the two factors extracted in principal components factor analysis with Spring 2007 Self-Concept and Locus of Control data: School year 2006–07	6-9
6-9	Varimax rotated factor patterns for the two factors extracted in principal components factor analysis with Self-Concept and Locus of Control item data: School year 2006–07	6-9
6-10	Self-Concept and Locus of Control weighted means and standard deviations, spring-eighth grade: School year 2006–07	6-10
6-11	Score breakdown, Self-Concept, by eighth-graders, sixth- and seventh- graders, and population subgroup: School year 2006–07	6-11
6-12	Score breakdown, Locus of Control, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006–07	6-12
7-1	Teacher rating scale reliability statistics for the IRT-based score, spring- eighth grade: School year 2006–07	7-4
7-2	Teacher rating scale means and standard deviations, spring-eighth grade: School year 2006–07	7-4
7-3	Percent of sample with perfect and minimum teacher rating scores, spring-eighth grade: School year 2006–07	7-:
7-4	English Oral Expression item difficulties (arranged in order of difficulty), spring-eighth grade: School year 2006–07	7-:
7-5	English Writing Skills item difficulties (arranged in order of difficulty), spring-eighth grade: School year 2006–07	7-0

<u>Table</u>		Page Page
7-6	Mathematics Skills item difficulties (arranged in order of difficulty), spring-eighth grade: School year 2006–07	7-6
7-7	Science Skills item difficulties (arranged in order of difficulty), spring- eighth grade: School year 2006–07	7-6
7-8	English Oral Expression standard errors, spring-eighth grade: School	7_7
7-9	English Writing Skills standard errors, spring-eighth grade: School year 2006–07.	7-7
7-10	Mathematics Skills standard errors, spring-eighth grade: School year 2006–07	7-8
7-11	Science Skills standard errors, spring-eighth grade: School year 2006–07	7-8
7-12	Intercorrelations among the indirect cognitive teacher ratings (ARS), selected child self-ratings (SDQ, Locus, Concept), and direct cognitive test scores, spring-eighth grade: School year 2006–07	7-10
7-13	Score breakdown, English oral expression, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006–07	7-12
7-14	Score breakdown, English writing skills, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006–07	7-13
7-15	Score breakdown, mathematics skills, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006–07	7-14
7-16	Score breakdown, science skills, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006–07	7-15
	Appendix A Tables	
A1	Reading assessment, unweighted sample sizes: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	A-1
A2	Mathematics assessment, unweighted sample sizes: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	A-2

<u>able</u>	
A3	Science assessment, unweighted sample sizes: School years 2001–02, 2003–04, and 2006–07
A4	Reading routing test number right, eighth grade assessment (range of possible values: 0 to 10): School year 2006–07
A5	Mathematics routing test number right, eighth grade assessment (range of possible values: 0 to 10): School year 2006–07
A6	Science routing test number right, eighth grade assessment (range of possible values: 0 to 10): School year 2006–07
A7	Reading IRT scale score, K-8 scale (range of possible values: 0 to 212): School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
A8	Mathematics IRT scale score, K-8 scale (range of possible values: 0 to 174): School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
A9	Science IRT scale score, 5-8 scale (range of possible values: 0 to 111): School years 2001–02, 2003–04, and 2006–07
A10	Reading T-scores, standardized within round (range of possible values: 0 to 96): School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
A11	Mathematics T-scores, standardized within round (range of possible values: 0 to 96): School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
A12	Science T-scores, standardized within round (range of possible values: 0 to 96): School years 2001–02, 2003–04, and 2006–07
A13	Reading IRT theta score, K-8 scale (range of possible values: -5 to 5): School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
A14	Mathematics IRT theta score, K-8 scale (range of possible values: -5 to 5): School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

<u>Table</u>		Page
A15	Science IRT theta score, 5-8 scale (range of possible values: -5 to 5): School years 2001–02, 2003–04, and 2006–07	A-15
A16	Probability of proficiency, reading level 1: letter recognition (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-16
A17	Probability of proficiency, reading level 2: beginning sounds (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-17
A18	Probability of proficiency, reading level 3: ending sounds (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-18
A19	Probability of proficiency, reading level 4: sight words (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-19
A20	Probability of proficiency, reading level 5: words in context (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-20
A21	Probability of proficiency, reading level 6: literal inference (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-21
A22	Probability of proficiency, reading level 7: extrapolation (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-22
A23	Probability of proficiency, reading level 8: evaluation (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	A-23
A24	Probability of proficiency, reading level 9: evaluating nonfiction (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-24

<u>Table</u>		Page
A25	Probability of proficiency, reading level 10: evaluating complex syntax (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-25
A26	Probability of proficiency, mathematics level 1: number and shape (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-26
A27	Probability of proficiency, mathematics level 2: relative size (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-27
A28	Probability of proficiency, mathematics level 3: ordinality, sequence (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-28
A29	Probability of proficiency, mathematics level 4: addition/subtraction (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-29
A30	Probability of proficiency, mathematics level 5: multiplication/division (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-30
A31	Probability of proficiency, mathematics level 6: place value (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-31
A32	Probability of proficiency, mathematics level 7: rate and measurement (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-32
A33	Probability of proficiency, mathematics level 8: fractions (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-33
A34	Probability of proficiency, mathematics level 9: area and volume (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07	A-34

<u>able</u>	
A35	Percent of children at or above modal reading proficiency for each grade: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
A36	Percent of children at or above modal mathematics proficiency for each grade: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
	Appendix B Tables
B1	Reading assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
B2	Mathematics assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
B3	Science assessment IRT item parameters: School years 2001–02, 2003–04, and 2006–07
	Appendix C Tables
C1	Reading assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
C2	Mathematics assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
C3	Science assessment estimated proportion correct: School years 2001–02, 2003–04, and 2006–07
	Appendix D Tables
D1	Reading assessment difference between actual and estimated percent correct by round: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

<u>Table</u>		<u>Page</u>
D2	Mathematics assessment difference between actual and estimated percent correct by round: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07	D-9
D3	Science assessment difference between actual and estimated percent correct by round: School years 2001–02, 2003–04, and 2006–07	D-16
	List of Figures	
Figure		
3-1	Three-parameter IRT logistic function for a hypothetical test item	3-8
3-2	Three-parameter IRT logistic functions for seven hypothetical test items with different difficulty (b)	3-9
3-3	Three-parameter IRT logistic functions for two hypothetical test items with different discrimination (a)	3-10
	List of Exhibits	
<u>Exhibit</u>		
7-1	Eighth-grade indirect and direct cognitive and noncognitive measures, examined for evidence of convergent and discriminate validity: School year 2006–07	7-9

1. INTRODUCTION

This report documents the design, construction, and psychometric characteristics of the assessment instruments used in the spring 2007 data collection of the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K). The ECLS-K is sponsored by the U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

The ECLS-K was designed to assess the relationship between a child's academic and social development and a wide range of family, school, and community variables. Analysis of the cognitive and social skills assessment scores described in this report, along with contextual variables in the ECLS-K database collected from schools, parents, teachers, and children, provides a basis for policy-relevant examination of growth rates, school influences, and subgroup differences in achievement and growth.

While the ECLS-K spans kindergarten through eighth grade, this report documents the psychometric properties for the seventh round of data collection, in spring 2007, when approximately 89 percent of the sampled children were in eighth grade. Also included is a review of the salient features of the assessments used in kindergarten through fifth grade. Among these salient features are the selection and design of assessment instruments and selected psychometric characteristics.

The ECLS-K eighth-grade assessment instruments include direct and indirect measures of children's achievement and socioemotional status. Direct measures refer to scores based on children's responses to self-administered cognitive test items and socioemotional inventories. Indirect measures refer to scores based on ratings completed by others such as teachers and parents. The mode of administration changed from an individually administered, computer-based assessment in the prior rounds to group-administered, paper-and-pencil assessment in eighth grade (see section 2.1).

The direct cognitive assessments for eighth grade were designed to measure an individual child's knowledge at a given point in time, as well as that same child's academic growth in each subject on vertical score scales based on successive assessments. The score scales for reading and mathematics measure growth from fall-kindergarten through eighth grade, while the science score scales measure growth only in the third- through eighth-grade rounds.

The cognitive assessments were designed not only to make reliable normative comparisons with respect to status and growth, but also to provide criterion-referenced interpretations. That is, in the reading and mathematics content domains, criterion-referenced proficiency scores can be used to describe a given child's mastery of specific knowledge and skills that mark ascending critical points on the developmental growth curve. These multiple criterion-referenced levels serve two functions. First, they help interpret what a particular attained score level means in terms of what a child can or cannot do. Second, they are useful in measuring change at particular points along the score scale. They provide a means of evaluating the relationship of certain school processes to changes in mastery of specific skills.

The development of the direct cognitive batteries used in kindergarten through eighth grade was carried out in five steps:

- 1. A review of the psychometric properties and constructs measured by existing assessment instruments.
- 2. Test frameworks were developed for the domains and constructs considered relevant for each grade.
- 3. Item pools were developed that reflected the test specifications in step 2.
- 4. Field tests of the item pools were conducted to gather statistical and psychometric evidence to identify the appropriate items necessary to achieve the assessment goals.
- 5. The final test forms were assembled consistent with field test item statistics and the frameworks.

Chapter 2 of this report describes the objectives and design of the eighth-grade assessment instruments. Differences between the kindergarten-first grade (K-1), third-grade, fifth-grade, and eighth-grade assessment batteries are described. For the direct cognitive tests, chapter 2 includes selection of content domains, notes on frameworks, descriptions of field testing, and selection of test items. It describes the criterion-referenced subsets of items in the reading and mathematics tests that were used to mark proficiency levels in kindergarten through fifth grade and the extension of these levels in reading for eighth-grade skills. Chapter 2 also describes the evaluation of potential gaps in the longitudinal scale for the years in which data were not collected (second, fourth, sixth, and seventh grades), and the steps taken to avoid compromising the measurement of gains. For the indirect measures, chapter 2 describes the development and content of the instrument used by children to rate their own academic ability and interest, and their behavior. Chapter 3 contains a description of the quality control procedures applied to analysis of the assessment data, as well as an overview of item response theory (IRT) procedures used in

computing test scores and the differential item functioning (DIF) procedures used to detect problem items. Chapter 4 presents the psychometric characteristics of the direct cognitive tests given in eighth grade, and chapter 5 describes their role in longitudinal measurement. Chapter 6 describes the development and psychometric characteristics of the self-description questionnaire administered to sampled children, and chapter 7 presents the same information for the teacher indirect cognitive rating scale measures.

A national probability sample of about 22,000 children in about 800 public and 200 private schools was assessed at entry to kindergarten in fall 1998 (round 1). They were followed up in springkindergarten (round 2), fall- and spring-first grade (rounds 3 and 4, respectively), spring-third grade (round 5), spring-fifth grade (round 6), and spring-eighth grade (round 7). While all base-year respondents were eligible for the spring-first grade data collection, fall-first grade was limited to baseyear respondents in a 30 percent subsample of schools. The seventh round of data collection described in this report took place in spring 2007, when approximately 89 percent of the children were in eighth grade. The direct cognitive assessments were conducted in all seven rounds of data collection, while the indirect cognitive measures were collected from teachers in rounds 1, 2, 4, 5, 6, and 7 (fall- and springkindergarten, spring-first grade, spring-third grade, spring-fifth grade, and spring-eighth grade). The indirect socioemotional measures were collected from teachers in rounds 1, 2, 4, 5, and 6, and from parents in rounds 1, 2, and 4. In rounds 5, 6, and 7 children completed a direct socioemotional measure. More details on the sample design and data collection methods used in the ECLS-K can be found in the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks (NCES 2009–004) (Tourangeau, Nord et al. forthcoming).

Sample counts, completion rates, psychometric characteristics, and score statistics for the eighth-grade assessments are presented in chapter 4 (direct measures) and chapter 6 (indirect measures), with score breakdowns by sex, race/ethnicity, socioeconomic status, and school type presented in appendix A. Additional information about the sample design, the assessment instruments, and the collection of assessment data can be found in the ECLS-K electronic codebooks and data file user's manuals. Statistics presented in this report may differ slightly from those in the current data file user's manual. Tables in the user's manual are based on the panel sample, that is, children who participated in all seven rounds of data collection, with national estimates computed using the longitudinal panel weight (C1_7SC0). The emphasis in this report is on the psychometric characteristics of the tests at each round, so all children participating in each round are included, and the corresponding cross-sectional weights,

(C1CW0–C7CW0), are used for national estimates. Statistics that report characteristics of the tests rather than national estimates, such as reliabilities or floor and ceiling effects, are unweighted. Detailed information on the assessments used in the earlier rounds can be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Fifth Grade* (NCES 2006–036rev) (Pollack, Atkins-Burnett et al. 2005).

2. DESIGN AND DEVELOPMENT OF THE ASSESSMENT INSTRUMENTS

The Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) assessment instruments were designed to measure children's academic and social development during the kindergarten through eighth-grade years. Direct and indirect cognitive measures describe children's academic performance at each time point, as well as measure growth over time. Measures of children's social behaviors were obtained through children's self-reports starting in the third-grade data collection. This chapter documents the design and development of the assessment measures used in the seventh round of data collection, when most of the ECLS-K children were in eighth grade.

The National Center for Education Statistics (NCES) and contractor staff assembled school curriculum specialists, teachers, and academics to consult on the design and development of the assessment instruments. Issues that were addressed included domains to be covered, test specifications, individual item content and presentation, mode of assessments, and time allocation. The advice of these experts guided the decisions necessary to ensure valid representation of domain content and to make efficient use of resources while minimizing burden on teachers and children.

The eighth-grade direct cognitive assessments built on the structure established in the kindergarten through fifth-grade rounds of data collection, with a change in the mode of administration, from computer-based to pencil-and-paper. Another design change in eighth grade from previous rounds was a reduction in the number of second-stage forms in each domain from three to two, with routing to only a low or high second-stage form, eliminating the middle form (discussed further below in section 2.1).

The content and components of the prior fifth-grade instruments were essentially similar to those used in third grade, with grade-appropriate increments in the difficulty of test items. The third-grade assessment battery differed from that of kindergarten and first grade (K-1) in several important respects. The English language screening assessment, assessment components of the parent questionnaire, and psychomotor assessment used in kindergarten and/or first grade were not included in the third-grade assessment battery. A questionnaire eliciting children's academic and behavioral self-ratings was added in third grade, and a science assessment replaced the K-1 general knowledge test.

Important changes in the assessments during the course of the longitudinal study are described here:

- No English language screening: In kindergarten and first grade, children who were identified as coming from a language minority background were administered an English language screening assessment, the Oral Language Development Scale (OLDS), prior to administration of the direct cognitive assessments. Once each child achieved a score sufficient for assessment in English, the OLDS was not administered to that child in subsequent rounds of data collection. At kindergarten entry, about 15 percent of the ECLS-K participants were found to need screening for English proficiency. By spring of first grade, less than 6 percent of the sample was screened, and nearly two-thirds of the screened children achieved the score required to go on to the rest of the assessment in English. Since no freshening of the sample occurred after first grade, the number of sampled children who might still lack English proficiency in third and subsequent grades was assumed to be so small that the language screening assessment would be unnecessary. Therefore, an English language screener was not administered after spring-first grade.
- No parent questionnaire items on children's social behaviors: Parents' ratings of children's behavior and social skills had been collected during the kindergarten and first-grade rounds. Parent ratings were discontinued after first grade for several reasons: age appropriateness of the items, technical issues (low intercorrelations among parent scales), and the need to minimize burden on participants.
- No psychomotor assessment: The fall-kindergarten assessment battery included an evaluation of children's fine and gross motor skills. This assessment was designed for use only in fall-kindergarten and was not repeated in subsequent rounds of data collection.
- Age-appropriate changes were made to the rating items used to measure children's perceptions of social skills and interest in school subjects. In the kindergarten and first-grade rounds of the ECLS-K, parents and teachers reported on children's social skills. In the third and fifth grade of the ECLS-K, the children provided information about themselves by completing a short self-description questionnaire that included items from a published instrument appropriate for third-and fifth-graders (Self Description Questionnaire I) (Marsh 1992a). In eighth grade, a new version of the self-description questionnaire was developed using items from a published instrument designed to be used with adolescents (Self Description Questionnaire II) (Marsh 1992b). In addition, two scales from the student questionnaire adapted from the National Education Longitudinal Study (NELS) measured children's self-concept and their perceptions of how much control they had over their own lives. See chapter 6 for more information on these scales and the scores that are available for analysis.
- Changes in the content and format of the direct cognitive assessment instruments: A change from the fifth-grade assessment design was a reduction in the number of second-stage forms in each domain from 3 to 2, with routing to only a low

or high second-stage form. This decision was based on several reasons, mainly due to the change from a computer-based, individual administration to a paper-and-pencilbased, group administration, from fifth grade to eighth grade. (See section 2.1 for details.)

- Longitudinal measurement: Similar to previous rounds, a portion of the prior round reading, mathematics, and science assessment items were included in the assessment for continuity and in anticipation of measurement of longitudinal gains. In earlier rounds, a science assessment, begun in third grade, replaced the direct cognitive assessment of general knowledge that had been used in kindergarten and first grade. A Spanish translation of the mathematics assessment, used in kindergarten and first grade, was assumed to be unnecessary for third, fifth, and eighth grades.¹ Details of these changes are described in section 2.1.
- Changes in the indirect cognitive assessment instruments: Separate teacher ratings of science and social studies skills in third grade replaced the K-1 general knowledge ratings. In fifth grade, the social studies section was discontinued in order to reduce teacher burden. In eighth grade, English, mathematics, and science teachers were asked to rate children on their respective domain-relevant skills (i.e., oral expression and writing skills, mathematics skills, and science skills, respectively). Information on the scaling of these items can be found in chapter 7.
- Elimination of data collection rounds: Another change in the original longitudinal design of ECLS-K was the elimination of the second- and fourth-grade rounds of data collection due to budgetary constraints. The implications of this decision, and the steps taken to minimize its impact on longitudinal measurement, are discussed in sections 2.1.5 and 5.1 of the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062) (Pollack, Rock et al. 2005). ECLS-K was originally planned to end with the fifth-grade data collection. The decision to add the eighth-grade round was made later. The impact on the longitudinal scale of not collecting data in sixth and seventh grades is discussed later in this report (section 2.1.3.2)

2.1 Direct Cognitive Assessment

The child development and education experts consulted by project staff during the design phase of each round of the ECLS-K recommended that the knowledge and skills assessed by each round of the ECLS-K tests should represent the typical and important cognitive goals of schools' curricula. Therefore, the subject-matter domains of language and literacy skills (referred to hereinafter simply as "reading" for the direct cognitive assessment), mathematics, and science were selected for the eighthgrade direct cognitive battery. Time constraints and concern with burden on children, as well as

¹ For more details on the Spanish mathematics assessment, see the *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99* (*ECLS-K*), *Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05) (Rock and Pollack 2002).

differences in social studies curricula throughout the states, led to a decision not to include a social studies assessment in the direct cognitive battery. (Although differences in science curricula exist throughout the states as well, it was decided to retain science for the direct cognitive assessments in order to measure gain.) The practical difficulties of adequately assessing children's proficiencies in writing, art, and music within the resource constraints of the study precluded assessment in these domains.

The nature of the ECLS-K cognitive assessment battery was shaped by its basic objectives and constraints. Foremost among these was the requirement that the test battery accurately measure children's cognitive development in reading and mathematics throughout the whole span of the study, and in science between third and eighth grades. The longitudinal design of the study required the development of vertical scales in each subject to support the calculation of valid change scores. Such scales would allow comparisons of achievement levels across grades and support estimates of the gains children make from year to year. The goal of minimizing time and burden on children and teachers determined the kinds of test items that could be used, as well as the structure of the tests. Some compromises were necessary to reconcile the goal of using age-appropriate reading passages with the objective of limiting total test time to an average of 80 minutes in eighth grade. The time limitation precluded the use of assessment tasks such as extended reading materials or hands-on science experiments.

As noted earlier, the same reading, mathematics, and general knowledge assessment instruments had been used in all four kindergarten and first-grade rounds of data collection. Children were routed to different levels of difficulty within each assessment domain depending on their performance on a short routing test in each subject area. For most children, the easiest of two (general knowledge) or three (reading and mathematics) second-stage forms was selected in fall-kindergarten, while by spring of first grade the majority of children were routed to the more difficult forms within the same sets. Because children's academic skills in third, fifth, and eighth grades could be expected to have advanced beyond the levels covered by the prior assessments, new sets of assessment instruments were developed for each round after those for the first grade. Some test items were retained from each round to the next to support development of a longitudinal score scale.

The K-1 general knowledge assessment, which included basic natural science concepts as well as concepts in social studies, was replaced by a science assessment administered in the third, fifth, and eighth grades. The science assessment is not comparable to the K-1 general knowledge assessment, so the longitudinal scale in science spans only the last three rounds of data collection. As a result, gains in

science can be measured only for third to eighth grade, while general knowledge scores can be compared only between the kindergarten and first-grade rounds.

Unlike the previous rounds of one-on-one computer-assisted assessments, the eighth-grade assessments were group-administered, paper and pencil tests. In addition, the assessments consisted of a routing test and two second stage forms, low and high, in each domain, rather than the three levels employed in the prior rounds. The routing tests were limited to 10 items in each domain to ensure that the field assessor could quickly score the routing tests on-site and issue the correct second-level tests.

The formats of the earlier assessments were similar, with some changes to accommodate the more advanced level of the questions. In the earlier years, an assessor presented the questions to the child and entered responses into a computer for each individually administered assessment. To accommodate the length of the reading materials used in the fifth-grade assessment, a separate booklet containing both the reading passages and questions was given to the child, with the questions also appearing on the easel handled by the assessor. In eighth grade, routing booklets for all three subjects were administered, and the children responded on answer sheets. While the assessor scored the routing form responses, the children completed the self-questionnaire. The second-stage booklets were then administered, based on the scores on the routing tests, with the children responding directly into the booklets.

The types of scores reported for the eighth-grade direct cognitive assessments are similar to those for kindergarten through fifth grade, with some modifications for scores representing both broadbased and criterion-referenced skills. Assessment scores from prior rounds were recalibrated and rescaled for eighth grade; the pool of items on which the broad-based scores are estimated was expanded to provide longitudinal measurement of gains in reading and mathematics for kindergarten through eighth grade, and in science for third to eighth grade. Thus, scores in the data files (both cross-sectional and longitudinal) for the earlier rounds should not be compared with recalibrated/rescaled scores from the kindergarten through eighth-grade round. Scores from the earlier rounds that are required for longitudinal measurement (e.g., scale scores and proficiency probabilities) have been rescaled and appear in the kindergarten through eighth-grade file in a metric that makes comparisons possible. A new reading proficiency level is defined in eighth grade that corresponds to the grade-appropriate skill of evaluating complex syntax. Descriptions of scores appear in chapter 4, and section 5.1.2 describes the procedures used to evaluate common functioning of items across different assessment rounds.

2.1.1 Administration of Adaptive Tests

During the design phase of the ECLS-K, project staff, which included experts in child development, primary education, and testing methodology, concluded that the ECLS-K would use individually administered assessments to sampled children. Since young children are not experienced test takers, individual administration could provide more sensitivity to each child's needs than a group-administered test. In addition to being individually administered, it was also recommended that the tests be adaptive in nature; that is, each child should be tested with a set of items that would be most appropriate for his or her level of achievement. The adaptive design of the assessments was continued through eighth grade, but with the change to a paper-and-pencil, group-administered assessment, which was more appropriate for children of middle school age.

The development of a vertical scale that must span kindergarten to eighth grade and have optimal measurement properties throughout the achievement range calls for multiple test forms that vary in their difficulty. The total pool of assessment items in each grade should reflect core curriculum elements for that grade. Within each grade, multiple test forms of varying difficulty optimize the accuracy of measurement for individuals with different levels of achievement. Overlapping items for forms within a grade, as well as across grades, link the forms to a vertical scale for measurement of longitudinal gains.

A child who is performing essentially on grade level should receive items that span the curriculum for his or her grade. A child whose achievement is above or below grade level should be given tasks whose difficulty level matches his or her individual level of development at the time of testing, rather than a grade-level standard. A child who is performing much better in relation to his or her peers, as measured by a brief routing test, would subsequently be given a second-stage form containing test items that are proportionately more difficult, while a child performing below grade level would receive a form with proportionately more easy items. The matching of the difficulties of the item tasks to each child's level of development that can take place in individualized adaptive testing situations increases the likelihood that the child will be neither frustrated by item tasks that are much too hard, nor bored by questions that are much too easy.

Psychometrically, adaptive tests are significantly more efficient than "one form fits all" administrations since the reliability per unit of testing time is greater (Lord 1980). Adaptive testing also minimizes 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. Adaptive testing uses performance at the beginning of a testing session to direct the selection of later tasks at an appropriate difficulty level for each child. Adaptive testing relies on item response theory (IRT) assumptions in order to place children who have taken different test forms on the same vertical score scale. Additional discussions of IRT may be found in chapter 3, and on the ECLS-K longitudinal scales in chapter 5.

A review of commercially available tests indicated that there were no "off-the-shelf" tests that matched the domain requirements and were adaptive and, for the early rounds, individually administered. Individual administration of assessments was considered important in the early years and was retained through fifth grade. The success of the adaptive approach in earlier rounds in optimizing measurement characteristics for a diverse sample of children suggested its use in the later grades as well.

2.1.2 The ECLS-K Frameworks

The ECLS-K was charged with assessing cognitive skills that are both typically taught and developmentally important. Neither typicality nor importance is easily determined. Identifying typical curriculum objectives and their relative importance is difficult because of the decentralized control that characterizes the American education system. The difficulties are compounded for the ECLS-K, since curriculum is constantly evolving and the data collection started with the kindergarten year in 1998, 2 years after the design phase, and continued until 2007.

For eighth grade, the National Assessment of Educational Progress (NAEP) content and process frameworks for reading, mathematics, and science were used as the basis of the assessment design for the ECLS-K round 7 data collection. The NAEP assessment goals are similar to those of the ECLS-K in that both projects aim to assess cognitive skills that schools typically emphasize. The NAEP 1992–2007 frameworks were particularly useful as models for the eighth-grade ECLS-K assessments since they define appropriate sets of skills and understandings at eighth grade. The resulting ECLS-K frameworks are fundamentally the same as the NAEP eighth-grade frameworks, with some differences due to ECLS-K formatting and administration constraints.

The NAEP frameworks are based on both current curricula and recommendations for curriculum change that have strong professional backing among theorists and teacher associations. NAEP is interested in the recommendations because it is charged with assessing skills and knowledge that reflect research in each domain and do not advocate a particular approach to instruction, but rather focus on important, measurable indicators of student achievement. These recommendations represent reasonable predictions about the directions that schools and school systems in the United States are likely to take in the near future and are thus appropriate to the ECLS-K. With respect to current curricula, NAEP relies on advice from panels of curriculum specialists. In addition to often being directly involved in the construction of curricula used in the schools, specialists often hold a wealth of local knowledge about current practices that is not recorded in publications and thus not otherwise available.

Despite these strengths, the NAEP test specifications have some important limitations in their applicability to the ECLS-K. NAEP frameworks define a number of different subscales within subject-matter domains, but test-length constraints forced the ECLS-K to define single proficiency scales for each subject domain. NAEP can measure multiple subscores within a content domain because it administers a large number of different item sets in a spiraled design to children at a given grade level. That design follows from NAEP's primary goal of measuring cognitive status at the *aggregate* level on a *cross-sectional* basis. In contrast, the ECLS-K attempts to attain relatively accurate *longitudinal* measurement (through adaptive test instrumentation and vertical scaling) at the *individual* level within a more focused cognitive domain.

In addition to the conceptual framework identifying the various types of skills and knowledge tested in the ECLS-K, the relative emphasis given to different content strands was designed to reflect typical curriculum emphases. The general rule used in determining allocations is that the composition of the tests should reflect typical curriculum emphases while considering differences in the number of items and length of items needed to adequately measure a given skill, knowledge, or concept. Systematically collected evidence on typical curriculum specialists and people with extensive teaching and administrative experience in schools and on the standards published by states and national professional organizations. For eighth grade, the overall testing time for each child was designed to consist of more time allotted for reading (due to the reading passages), with a lesser amount of time allocated for the mathematics and science assessments. It is important to keep in mind that some content strands can be assessed more quickly than other areas. For example, many mathematical computation items can be

administered in a short period of time, while reading questions based on passage comprehension require a greater investment of time.

Tables 2-1 to 2-3 present the test specifications for the ECLS-K cognitive battery from kindergarten through eighth grade. The numbers in the cells are the target percentages for each content area; they are at best approximations since the item classifications are somewhat arbitrary. Particularly in third, fifth, and eighth grades, many items measure more than one area. For example, solving a mathematics problem may require understanding of number concepts as well as skill in interpreting data. The items for the kindergarten and first grade were allocated according to the amount of time items were expected to take. However, for the third, fifth, and eighth grades, the content allocations were based on counts of numbers of items matching the content frameworks as closely as possible.

2.1.2.1 Reading Test Specifications

The ECLS-K reading specifications were adapted from the 1992 and 1994 NAEP Reading Frameworks (National Assessment Governing Board [NAGB] 1994a) for the early rounds, and from the 1992–2007 frameworks for eighth grade. The NAEP framework is defined in terms of four types of reading comprehension skills:

- **Initial understanding** requires readers to consider the text as a whole and provide a global understanding of it. Explaining the purpose of an article, reflecting on the theme of a story, or identifying the topic of a passage would be included in this category.
- **Developing interpretation** requires readers to develop a more complete understanding of what was read. It involves focusing on specific information in the text as well as linking information across parts of the text. Testing the meaning of vocabulary words in the text would be included in this category.
- **Personal reflection and response** requires readers to connect information in the text with their background knowledge and experience in the real world. Supporting an opinion about an issue raised in a historical text with examples from contemporary life would be included in this category.
- **Demonstrating a critical stance** requires readers to stand apart from the text, consider it objectively, and judge its appropriateness and quality. Evaluating language and textual elements, thinking about the author's purpose and style, and making connections between two texts would be included in this category.

The NAEP frameworks are defined for fourth, eighth, and twelfth grades; therefore, the eighth-grade frameworks were directly applicable to the ECLS-K round 7 data collection. However, the NAEP fourth-grade frameworks had to be modified for the earlier rounds of the ECLS-K to accommodate adequately the basic skills typically emphasized beginning in kindergarten through fifth grade. In the kindergarten and first-grade rounds, two skill categories were added to the NAEP framework: Basic Skills, which includes familiarity with print, recognition of letters and phonemes, and decoding, and Vocabulary. After first grade, the emphasis on basic skills in the ECLS-K reading framework was decreased, so that the allocations for third and fifth grades are very close to that of the reading comprehension skills of fourth grade NAEP. Literacy curriculum specialists and teachers contributed to development of the framework and reviewed item pools. The conceptual categories shown in table 2-1 combine the recommendations of the literacy curriculum specialists with the NAEP reading framework.

Notably absent from the ECLS-K reading framework is any place for writing skills. This absence is a reflection of practical constraints associated with limited amount of testing time and the cost of scoring. Nevertheless, the ECLS-K asked teachers to provide information on each sampled child's writing abilities with the use of the Academic Rating Scale (see chapter 7 in this report).

				Re	ading comprehension skills	
Grade levels	Total	Basic Skills	Vocabulary	Initial understanding/ developing interpretation	Personal reflection	Critical stance
			Percent	of testing time		
Kindergarten	100	40	10	35	10	5
First grade	100	40	10	35	10	S
			Percen	t of test items		
Third grade	100	15	10	45	15	15
Fifth grade	100	10	10	45	15	20
Eighth grade	100	0	0	55	15	30
NOTE: The content strands i familiarity with print, recogni were separated in prior rounds	rre identical to the 1 tion of letters and ph	Vational Assessment of Ec nonemes, and decoding. Th	lucational Progress Reading te framework categories of i	Framework categories, wi nitial understanding and dev	th the addition of Basic Skills and Vo eloping interpretation were combined	cabulary. Basic Skills include in the eighth-grade design, but

Reading longitudinal test specifications for kindergarten through eighth grade: School years 1998–99, 1999–2000, 2001–02, 2003–04 and 2006–07 Table 2-1.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2000, spring 2002, spring 2004, and spring 2007.

2.1.2.2 Mathematics Test Specifications

The mathematics test specifications shown in table 2-2 are primarily based on the Mathematics Framework for the 1996 NAEP (National Assessment Governing Board [NAGB] 1996a) modified for the rounds prior to eighth grade, and on the 2005 NAEP Mathematics Framework for the eighth grade, which in turn were derived from the curriculum standards from the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics (NCTM). The content strands represented by the column categories in table 2-2 are defined as follows:

- Number sense, properties, and operations. In eighth grade, this content area largely assesses number sense. Number sense is defined as comfort in dealing with numbers effectively. It includes firm intuitions about what numbers tell us; an understanding of the ways to represent them symbolically (including facility with converting between different representations); the ability to calculate, either exactly or approximately; and skill in estimation. The ability to deal with proportion, including percents, is another important part of number sense. In eighth grade, children should be proficient with rational numbers, represented either as decimal fractions (including percents) or as common fractions. They should be able to use them to solve problems involving proportionality and rates. In middle school also, numbers should begin to coalesce with geometry via the idea of the number line. This should be connected with ideas of approximation and the use of scientific notation. These children should also have some acquaintance with naturally occurring irrational numbers, such as square roots and pi.
- Measurement. Measuring is the process by which numbers are assigned in order to describe the world quantitatively. This process involves selecting the attribute of the object or event to be measured, comparing this attribute to a unit, and reporting the number of units. Attributes such as capacity, weight/mass, time, and temperature are included, as well as the geometric attributes of length, area, and volume. Units involved in items on the assessment include nonstandard, customary, and metric units. Eighth grade includes the use of both square and cubic units for measuring area, surface area, and volume; degrees for measuring angles; and constructed units such as miles per hour.
- Geometry and spatial sense. In this content area, children 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 eighth grade, questions about cross-sections of solids and the beginnings of an analytical understanding of properties of plane figures, especially parallelism, perpendicularity, and angle relations in polygons are included. Right angles and the Pythagorean Theorem are introduced, and moving toward the high school level, geometry becomes more and more mixed with measurement. Questions on symmetry and transformations are also a part of this content area.
- Data analysis, statistics, and probability. Data analysis covers the entire process of collecting, organizing, summarizing, and interpreting data. In the context of data analysis, or statistics, probability can be thought of as the study of potential patterns in

outcomes that have not yet been observed. In eighth grade, children are expected to be able to describe distributions of data through center, spread, and shape, use a variety of organizing and summarizing techniques, and begin to use formal terminology related to probability and data analysis.

Patterns, algebra, and functions. In eighth grade, central topics in this content area include assessing the ideas of function and variable. Representation of functions as patterns, via tables, verbal descriptions, symbolic descriptions, and graphs, can combine to promote a flexible grasp of the idea of function. Linear functions receive special attention. They connect to the ideas of proportionality and rate, forming a bridge that will eventually link arithmetic to calculus. Symbolic manipulation in the relatively simple context of linear equations is reinforced by other means of finding solutions, including graphing.

				Content strands		
Grade levels	Total	Number sense, properties, and operations	Measurement	Geometry and spatial sense	Data analysis, statistics, and probability	Patterns, algebra, and functions
		Percent o	f testing time			
Kindergarten	100	50	15	5	10	20
First grade	100	50	14	10	10	16
		Percent	of test items			
Third grade	100	40	20	15	10	15
Fifth grade	100	40	20	15	10	15
Eighth grade	100	20	15	20	15	30
NOTE: The content strands are identical to the Nation SOURCE: U.S. Department of Education, National C spring 2000, spring 2002, spring 2004, and spring 200	al Assessment of Center for Educati)7.	Educational Progress Math on Statistics, Early Childh	ematics Framework catego ood Longitudinal Study, K	ries. Lindergarten Class of 1998.	–99 (ECLS-K), fall 1998, sp	ring 1999, fall 1999,

Mathematics longitudinal test specifications for kindergarten through eighth grade: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07 Table 2-2.

2.1.2.3 Science Test Specifications

The K-1 general knowledge test, a combination of science and social studies items, was replaced by a science test for third, fifth, and eighth grades. No direct measurement of social studies knowledge was included in third through eighth grades, although teacher ratings of children's proficiency in social studies were collected in third (but not fifth or eighth) grade. For a discussion of the design and specifications of the K-1 general knowledge test, refer to the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05) (Rock and Pollack 2002).

The test specifications for eighth-grade science (table 2-3) were derived from information in the 1996 NAEP Science Framework (National Assessment Government Board [NAGB] 1996b) and the results of a NAEP 2000 survey on science course-taking patterns by middle and high school students. The test specifications for third- and fifth-grade science were developed largely from recommendations of the ECLS-K advisory group. Similar to the 1996 NAEP Science Framework, the ECLS-K science framework includes two broad classes of science competencies: conceptual understanding and scientific investigation.

- **Conceptual understanding** refers to both the child's factual knowledge base and the conceptual accounts that children have developed for why things occur as they do. Consistent with current curriculum trends, the emphasis in the ECLS-K is more on the adequacy of accounts than the grasp of discrete facts, particularly as the children moved up in grade level.
- Scientific investigation refers to children's abilities to formulate questions about the natural world, to go about trying to answer them on the basis of the tools available and the evidence collected, and to communicate their answers and how they obtained them.

The ECLS-K science assessment includes questions drawn from the fields of earth, physical, and life science. These fields are defined as follows:

• Earth and space science is the study of Earth's structure and systems as well as its place in the universe. Children in eighth grade are expected to know the nature of the layers of the solid Earth and the related dynamic processes that cause it to change, such as the rock cycle and the movement of tectonic plates. Children should have knowledge of the water systems and atmospheric systems and be able to describe how these systems interact causing the water cycle. They should also have an understanding of how the relative motions of the components of the solar system
cause day and night, the seasons, eclipses, etc., and should be able to describe the formation of the solar system.

- Physical science includes matter and its transformations, energy and its transformations, and the motion of light, sound, and physical objects. Children should have an understanding that matter is composed of atoms and molecules on a microscopic scale, and be able to classify materials into elements, compounds, and mixtures. The understanding of the conservation of matter, as well as the properties of matter such as conductivity and solubility, are assessed. Physical and chemical changes are assessed in molecular terms. Children should be able to recognize energy in its various forms, and be able to describe energy transformations and consequences of the conservation of energy in both natural and human-made systems. Both qualitative and quantitative aspects of motion of objects on macroscopic scales are assessed in terms of distance, speed, time, and Newton's Laws that describe the results of forces on objects. The characteristics and motion of waves, both sound and light, are part of this field.
- Life science topics include cells and their functions, organisms, diversity, and ecology. Beginning at the middle school level, diversity includes an understanding of genetic variations within species and theories of adaptation and natural selection. Organisms should be understood in terms of their reproduction, growth, and development, life cycles, and functions and interactions of body systems within organisms. Ecology addresses the interactions of organisms with their environment, both living and nonliving. Included in ecology is the flow of energy into and through organisms and ultimately through the ecosystem, an understanding of factors that cause changes in populations, and the environmental effects of human activity.
- Table 2-3.Science longitudinal test specifications, in percent of test items, for third grade (spring 2002),
fifth grade (spring 2004), and eighth grade (spring 2007)

		Earth and space		Life
Grade levels	Total	science	Physical science	science
Third grade	100	33	33	33
Fifth grade	100	33	33	33
Eighth grade	100	40	30	30

NOTE: The ECLS-K science expert panel developed the content strands and target allocations. The allocation of items at each grade level follows the 1996 NAEP guidelines that specify that about half of the items within each of the science subdomains measure conceptual understanding and half measure scientific investigation. Detail may not sum to totals due to rounding.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002, spring 2004, and spring 2007.

In terms of subject matter emphasis in the elementary grades, the 1996 NAEP Science Framework, American Association for the Advancement of Science (1995) and National Academy of Sciences (1995) recommend roughly equal emphasis on the three strands: earth, life, and physical science. Review of elementary text series (Harcourt Brace 1995; Holt 1986; Scott-Foresman 1994; and Silver

Burdett & Ginn 1991) revealed that coverage of these topics is equally distributed. The ECLS-K advisors concurred with the recommendation of equal representation of the strands at each grade level, and the final item batteries reflect that balance. For eighth grade, an increased emphasis on life science is consistent with the NAEP frameworks for that grade.

2.1.3 Field Testing of Direct Cognitive Items

Prior to the national data collection a field test was conducted to evaluate the psychometric properties of the direct cognitive items, and a pilot test was undertaken as well to ensure that the procedures for conducting the assessments worked smoothly and yielded accurate test scores. The field test was conducted in the spring of 2006. Relatively small samples of children participated in the field test, and relatively large numbers of test questions were tried out. Both multiple-choice and open-ended items were used in reading and mathematics, with multiple-choice only items in the science domain. Items were revised on the basis of the field test results, and sets of questions were selected for the full-scale test for eighth grade.

During the fall of 2006, the pilot test, as noted above, was conducted to evaluate the procedures for administering the assessments. The pilot test reviewed operational procedures to confirm the script and evaluate the flow of the entire assessment (i.e., routing form and scoring templates, student questionnaire, second-stage tests, and measurement of height and weight). The routing form items were examined only in terms of accuracy in correctly scoring those items and identifying and labeling the appropriate second-stage tests.

The remainder of this section reports on the field testing of the cognitive assessment.

2.1.3.1 Field Test Design

Both eighth- and tenth-graders were included in the spring 2006 field test sample, in anticipation of a tenth-grade national round of data collection, which was subsequently canceled. Thus field test results were used to guide the revision and selection of items for only eighth-grade assessments for the longitudinal sample.

Field test issues. The field test for eighth grade, as for earlier rounds, was designed primarily to gather the necessary psychometric data to evaluate the suitability of items for selection for the operational test forms. An additional purpose for the field test for the earlier grades was the construct validation of the reading and mathematics item pools, by comparison of field test results with scores on selected sections of an established assessment instrument. See the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Fifth Grade (NCES 2006– 036rev) (Pollack, Atkins-Burnett et al. 2005) for further details. A validation instrument was not administered during the eighth-grade field test. It has been common practice in the early childhood longitudinal studies to validate proposed item pools by administering an accepted off-the-shelf test to the field test sample and comparing results. This was not done for the eighth-grade field test since the design and content of the assessments were based on frameworks and test items from previously validated surveys, including the National Education Longitudinal Study of 1988 (NELS:88), Education Longitudinal Study of 2002 (ELS:2002), ECLS-K round 6 (fifth grade), and released National Assessment of Education Progress (NAEP) items, and was modeled on NAEP, which in turn had been validated previously. The resulting high correlations from the prior validations supported the validity of the source tests and thus can be interpreted as also supporting the ECLS-K eighth-grade item pool.

Spring 2006 Field Test. Cognitive test items in reading, mathematics, and science were administered in the Spring 2006 Field Test. A total of 95 unique items in reading, 100 in mathematics, and 65 in science were administered. Items in each subject area were distributed among multiple forms with approximately parallel content and difficulty. Two forms in mathematics and science and four forms in reading per grade were sorted into eight booklets, each containing one form in each of two subject areas, laid out so that each set of reading questions appeared as the first section in one booklet, and either mathematics or science as the second section. The eight booklets were spiraled among the approximately 3,600 eighth- and tenth-grade test takers participating in the field test. This resulted in approximately 400–800 observations for each test item, depending upon overlap on other forms within and across grades. Those items appropriate for both eighth and tenth grade were presented on multiple forms and resulted in more observations; others, occurring on only single forms, resulted in fewer observations. Table 2-4 shows the organization of the field test booklets.

Approximately 300 more tenth-grade than eighth-grade respondents participated in the field test. Results were analyzed for both grades combined since the emphasis was on evaluating the performance of the items across a broad range of ability levels and maintaining maximum sample sizes to

help stabilize estimates. For issues that relate directly to planning for the eighth-grade testing, such as the difficulty of the items, the focus was predominantly on the eighth-grade part of the sample only.

Booklet	Observations	Section 1	Section 2
Grade 8 Booklet 1	382	Reading Grade 8 Form 1	Mathematics Grade 8 Form 1
Grade 8 Booklet 2	378	Reading Grade 8 Form 2	Science Grade 8 Form 1
Grade 8 Booklet 3	379	Reading Grade 8 Form 3	Mathematics Grade 8 Form 2
Grade 8 Booklet 4	388	Reading Grade 8 Form 4	Science Grade 8 Form 2
Grade 10 Booklet 1	457	Reading Grade 10 Form 1	Mathematics Grade 10 Form 1
Grade 10 Booklet 2	466	Reading Grade 10 Form 2	Science Grade 10 Form 1
Grade 10 Booklet 3	461	Reading Grade 10 Form 3	Mathematics Grade 10 Form 2
Grade 10 Booklet 4	455	Reading Grade 10 Form 4	Science Grade 10 Form 2

Table 2-4.Number of observations from the ECLS-K field test pool in field test forms, by section:Spring 2006 field test

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 Eighth/Tenth Grade Field Test, spring 2006.

Each set of reading items appeared as the first section of one test booklet with either a mathematics or science form as the second section. It is possible that test performance might be improved by a practice effect, that is, a test taker performing better on items administered toward the end of a test with earlier items serving as practice tasks. Conversely, if a fatigue effect is operating, children may do better on items administered near the beginning, before they have become tired. It was determined in previous rounds of the ECLS-K that the practice and fatigue effects were negligible, and it was decided to not counterbalance the domains for the 2006 field test. In addition, since reading would be presented as the first domain in the national assessment forms, it was preferred to do the same in the field test.

Each of the eight reading field test forms had four reading passages and approximately 20 items, in a mix of multiple-choice and open-ended format. Several passages and associated items were presented on multiple forms within and across grades. Other passages and item sets were taken from the operational fifth-grade assessment. Items from the operational fifth-grade assessment were included in the field test in anticipation of inclusion in the eighth-grade national assessment. The overlap of items between fifth and eighth grade provided a strong link anchoring the scale for the purpose of measurement of gain.

The eighth- and tenth-grade field test contained 125 mathematics items, divided among four forms, with the two forms in each grade designed to be approximately parallel with respect to the content and difficulty of the items. Each form appeared in one test booklet, paired with a reading form. Some

items appeared in multiple forms within or across grades. Others were obtained from the fifth-grade operational assessments. As stated above, inclusion of fifth-grade items was in anticipation of selecting these items for the eighth-grade national assessment, for linking purposes for longitudinal measures of gain. Both multiple-choice and open-ended items were presented in each form.

Two eighth-grade and two tenth-grade field test science forms each contained 20 items. Each form within grade was designed to be parallel within content and item difficulty. Similar to the mathematics, each form appeared in one test booklet, paired with a reading form. Some items appeared on multiple forms, within or across grades. Some items from the fifth-grade operational assessments were retained in anticipation of scaling for longitudinal measurement. Only multiple-choice science items were presented on the field test. Response time for open-ended science items was estimated to be longer than that for the reading and mathematics items; therefore, in order to maximize the number of items presented within minimal time, open-ended items were not included in the field test forms.

2.1.3.2 Field Test Results and Conclusions

Analysis of field test data focused on both psychometric characteristics of the test items and operational issues. Psychometric analysis included calibration of item difficulty and discrimination, identification of flawed items, and detection of differential item functioning (DIF) with respect to population subgroups. Operational issues examined included timing, completion rates, and level of cooperation. Comprehensive reports from the assessors who administered the field tests complemented the analysis of item response data and played an important part in the design of the eighth-grade assessments.

Psychometric characteristics of test items. Classical item statistics were obtained for each of the field test items. Item difficulty was represented by percent correct, which was computed for eighthand tenth-grade participants combined, as well as for each grade separately. Item discrimination, that is, the extent to which each item is consistent with the overall set of items, was measured by *r*-biserials, which are correlations of total score with item score (right/wrong) for each item. Distractor analysis consisted of evaluating statistics on the percentage of children choosing each response option for multiple-choice items, and the average total test score for those choosing each option. This information provided a basis for identifying items that might have more than one potentially correct answer, items with incorrect response options chosen by children scoring higher, on average, than those choosing the

intended correct option, or items with response options that seemed so implausible that few if any children selected them. Item analysis procedures provided information on the number of children who omitted each item and their performance on the test as a whole. A high number of omitted items, for children who then went on to answer subsequent test questions, can be an indication that a test item is confusing or otherwise problematic for children. Classical item statistics also included the alpha coefficient, a measure of reliability, for each set of field test items.

IRT parameters (Lord 1980) were estimated for all cognitive items in the field test for the purpose of item selection only, using the PARSCALE computer program (see section 3.2.2 for details). (Parameters were re-estimated later using national sample data.) The IRT parameters were based on the three-parameter model with a parameter for guessing, a parameter for difficulty, and a slope (discrimination) parameter. The IRT slope, or "a" parameter, complements the information provided by the *r*-biserial but relates item discrimination to overall performance at a particular ability level rather than for the whole range of ability. The "b" parameter provides a measure of difficulty that is less susceptible to distortion, if large numbers of children omitted an item, than is percent correct. Marginal maximum likelihood estimation procedures (Mislevy and Bock 1982; Muraki and Bock 1991) were used to estimate the item parameters. Item characteristic curves (ICCs) were inspected for indications of lack of fit. Graphs containing the ICCs also included markers showing percent correct, separately for eighth- and tenth-graders, at intervals spaced along the ability range. This permitted evaluation of overall fit as well as displaying possible differences in functioning for the two grades. A relatively small percentage of items exhibited overall lack of fit and were removed from consideration for the eighth-grade battery.

IRT-based estimates of ability distributions provided a basis for the selection of target difficulty ranges for the eighth-grade test forms. The metric of the IRT ability estimates for field test participants corresponds to the metric of the item difficulty parameters. This allowed the selection of items with difficulty matched to the ability levels that could be expected in the eighth-grade assessment. Although the field test sample was not designed to be nationally representative, care was taken to select participating schools such that the sample would include both high and low achievers. Section 2.1.4 describes the use of the item difficulty and ability parameters in the selection of items for the eighth-grade forms.

The question of whether the absence of sixth- and seventh-grade rounds of data collection might result in a gap in ability levels that might seriously impact the measurement of gain was addressed. Examination of the field test results showed a considerable overlap in ability distributions between fifth-

and eighth-graders. As a result, no sixth- or seventh-grade "bridge" data collection, analogous to the second-grade sample that had been assessed to bridge the first-to-third grade gap, was necessary.

Cognitive test items were checked for DIF for males compared with females. There were too few Hispanic and Asian children in the field test sample for DIF analyses to be carried out for these groups. Sample sizes of Black children were sufficiently large for Black/White DIF to be evaluated for most of the field test items. It is not necessarily expected that different subgroups of children will have the same average performance on a set of items. But when children from different groups are *matched on overall ability*, performance on each test item should be about the same. There should be no relative advantage or disadvantage based on the child's gender or racial/ethnic group.

The DIF procedure (Holland and Thayer 1986) is designed to detect possible differential functioning for subgroups by comparing performance for a focal group (e.g., females or Black children) with a reference group (e.g., males or White children), while holding ability constant. DIF refers to the identification of individual items on which some population subgroups (the focal groups) perform, on average, relatively better or worse in comparison with members of a reference group who are matched in terms of overall performance on the total pool of items. Items are classified as "A," "B," or "C" depending on the statistical significance of subgroup differences, as well as effect sizes. Items identified as having "C" level DIF have detectable differences that are both sizeable and statistically significant. Chapter 3 provides a more detailed description of the procedures used to detect DIF levels of items.

A finding of differential functioning, however, does not automatically mean that a test item is inappropriate. It simply means that the item is differentially easier or more difficult for some subgroup (focal group) when compared with a reference group. A judgment that an item is inappropriate requires not only the statistical measure of DIF for one or more subgroups, but also a determination that the difference in performance is *irrelevant to the construct being measured*. In other words, different population subgroups may have differential exposure or skill in solving test items relating to a topic included in the test specifications. If so, the finding of differential performance may be an important and valid measure of the targeted skill and should be included in the assessment (see section 3.4; also Holland and Thayer 1986). Items that demonstrate differential functioning favoring the reference group were reviewed for inappropriate content by a standing committee on test fairness at Educational Testing Service (ETS), consisting of both majority and minority group members. Items that were judged to have content or presentation that might be problematic for a particular focal group in ways that are not relevant to the construct being measured were dropped from the item pool. Some items that had DIF that was judged to be the result of possible differential skills in some area of the test framework, and not merely due to subgroup membership, were retained. The remaining pool of items was sufficient for assembly of the eighth-grade national assessment forms.

Three reading items exhibited negative C-level DIF *against* the focal group for females (two items) and Black children (one item). One C-level DIF item *favoring* the focal group was found for each of these groups. All but one of these items were deleted from the eighth-grade assessment pool. One item favoring females was included on the eighth-grade routing form. This item was originally administered in NAEP and, based on a much larger sample, showed no evidence of DIF. And in the eighth-grade field test, the DIF statistics for this item were borderline C-level. This was assumed to be the result of instability due to the small sample sizes in the field test. It was recommended that this item be retained for eighth grade because it had good statistics and a difficulty level appropriate for the routing form.

In the mathematics field test, one multiple-choice item exhibited C-level DIF against females. This item was originally presented in the fifth-grade assessments in open-ended format. Early on in the field test design, a few items were selected to be presented in both multiple-choice and open-ended format. Discussions after the field test was complete resulted in discarding the modified items for use in design of the national forms, so DIF for this item is not relevant. Four other mathematics items exhibited C-level DIF against Black children and were not retained for the eighth-grade operational forms. Two items showing C-level DIF *favoring* females were selected, and another favoring Black children. In general it is recommended to remove items showing C-level DIF against the *focal* group (e.g., a minority group) from any subsequent assessments. Policy regarding C-level DIF against the *reference* group (the comparison group, e.g., White children) is not constrained in the same manner, as is the case in the proposed mathematics forms. One item favoring females was included on the eighth-grade operational forms for several reasons: it had good statistics, was a linking item from the fifth-grade round, and did not exhibit DIF in the fifth-grade round, so the DIF observed in the field test is assumed to be the result of instability in the estimate due to the small sample size.

For science, negative C-level DIF against females was detected for one item and two against Black children. These items were not retained for the eighth-grade operational forms. C-level DIF *favoring* females was found for one item, and favoring Black children for another. As with the mathematics forms, one item favoring the focal group (females) was recommended for the eighth-grade national forms for the same reasons. The item had good statistics, was a linking item from the fifth-grade round, and did not exhibit DIF in the fifth-grade round, when analysis was performed on a much larger sample size.

2.1.4 Eighth-Grade Test Forms

The eighth-grade assessments were designed to support measurement of the reading, mathematics, and science domains as accurately as possible, at all levels of ability found within the ECLS-K eighth-grade round and longitudinally as well. Assembly of the test forms from the field-tested items took into account numerous objectives, including psychometric considerations, framework specifications, and practical issues. The psychometric considerations included item quality and reliability, item difficulty, floor and ceiling effects, and longitudinal measurement. Field-tested items were candidates for selection for final test forms if they had acceptable item analysis statistics and IRT parameters and were not rejected due to DIF problems related to subgroup membership. Framework specifications, and practical issues such as timing and scorability of items, placed additional constraints on assessment design. Design of the test forms required some compromises due to competing objectives.

The final administration consisted of each child receiving a booklet containing routing tests in reading, mathematics, and science, consisting of 10 questions each, with responses written on a paper answer sheet. Upon completion of the routing tests, the children then completed the student questionnaire while, at the same time, the administrator scored the three routing tests using a scoring template and selected the second-stage forms indicated for each child/subject. The timed, second-stage forms were then administered after the student questionnaires were completed.

2.1.4.1 Item Quality and Reliability

To contribute useful information about children's skill levels, test items selected for the final forms should ideally have high *r*-biserials (0.4 or higher) and IRT "a" parameters (1.0 or higher), as well as good fits of empirical data to the IRT model. Items with high discrimination parameters permit accurate placement on the ability continuum. A small number of the selected items fell short of these standards but were selected for other reasons such as framework specifications, overlap with fifth-grade assessments, or links to a selected reading passage. In IRT, the measurement precision for individual examinees is improved by administering the maximum number of items possible in the time available,

and including items that function appropriately and measure the same construct. Items found to have DIF for population subgroups were deleted from the item pool except as noted earlier.

The design change from an individually administered computer-based assessment in kindergarten through fifth-grade to group-administered paper-and-pencil assessment in eighth grade required a change in the number of second-stage forms, as indicated above. The impact on reliability of using two second-stage forms instead of three was reviewed. Two types of reliabilities were examined, the reliability of the whole assessment score (reliability of theta), and the internal-consistency reliability (coefficient alpha) of each form separately.

At one extreme of the possible number of forms would be a single test form: the same test given to everyone, regardless of ability. Assuming that the test items are of appropriate difficulty for the sample (i.e., good variation in performance), the coefficient alpha would be relatively high, because the variance of scores for people taking the form is at a maximum (that is, the variance of the whole sample). Dividing the sample according to ability (adaptive testing) and assigning groups to harder or easier test forms means that the variance within each form is restricted, and thus the alpha reliability would be lower for separate forms than for a single form.

The reliability of the whole assessment works the other way. At the opposite extreme, with a very large number of test forms, such as a potentially different form for each person (as in computeradaptive tests), the result is a high reliability of theta (and the IRT-related scores) because accuracy of measurement is good for each person (minimizing floor and ceiling effects), but a very low alpha reliability for each "test form" because the variance of each one is very limited.

Because the sample taking the eighth-grade second-stage forms is divided into fewer groups, two instead of three, the variance *within* each form can be expected to be greater, and the coefficient alpha greater for each of the two forms than if there were more forms.

The issue that deserved attention was whether the reliability of the whole assessment (that is, of the theta ability estimate) was at risk. If fewer forms meant that each child's ability was less well estimated because the test items received were not of the right difficulty, the reliability of theta could have been affected. This would have been most likely to be observed in the tails of the ability distribution, where the lowest and highest achievers might not have received appropriate sets of items. To test this, simulations were run to estimate the abilities in the eighth-grade national sample. The results showed no

floor or ceiling effects for the design using two second-stage forms, even with the inclusion of estimates for children who would be below or above grade level. In addition, the distribution of simulated numberright on each form had a wide range, no "clumps," suggesting that children throughout the ability range would receive the items necessary for accurate measurement. This suggested that the reliability of theta for this design would continue to be high, similar to previous rounds in the ECLS-K.

2.1.4.2 Item Difficulty

Accurate measurement at all scale points requires that children receive sets of test items that are close to their ability level. The routing section of each assessment should direct each child to an appropriate set of second-stage items. Within each second-stage form, the item difficulties were selected to match the expected ability levels of the test takers. The distribution of IRT ability estimates for the field test eighth-graders was used to determine item difficulty objectives. The low and high second-stage forms emphasized easier and harder items, respectively. The test items taken by each child (routing test plus one second-stage form) were designed to have a rectangular distribution of item difficulties in the target ability range, that is, IRT "b" parameters that were approximately equally spaced with no large gaps.

2.1.4.3 Floor and Ceiling Effects

Floor effects occur when all test items are so difficult that many children must simply guess at random, while ceiling effects are a result of a test that is too easy, with many children achieving a perfect score. Tests that are too hard or too easy for large numbers of test takers do not do a good job of measuring the ability levels of the lowest and highest achieving children. It is particularly important to avoid floor and ceiling effects in a longitudinal study so that achievement gains may be measured accurately. The eighth-grade assessment forms were designed to have enough easy items that distinctions could be made at the low end of the ability range, and enough hard items to accurately measure the most skilled children. To avoid floor and ceiling effects, each assessment included a few items in the high second-stage form that almost all children would get wrong, and a few in the low second-stage form that almost all children would get right, so that accurate measurement of the extremes of ability could be accomplished. Each of the second-stage test forms contained some items with difficulty levels that extended beyond the target ability range, at both the high and low end. This design feature served two purposes. First, it provided some of the overlapping items required to put all of the test forms on a common scale (in addition to routing items taken by all children). Second, it improved measurement properties for children whose achievement level was very near a routing cut point. There was the possibility that guessing and/or careless mistakes on the routing test could result in children at the margin receiving a second-stage test form that was too easy or too hard. For example, a child whose ability level was near the low end of the upper ability range might miss a few routing test items and be assigned to the low second-stage form. Accuracy of measurement in this situation was supported by the overlap of some of the hardest low-form items with the easiest high-form items.

2.1.4.4 Longitudinal Score Scale

Measurement of gain over time requires a longitudinal score scale. The challenge for the ECLS-K was to establish a common scale not only for tests given in different grades but also for different forms of the test within each grade. In the four rounds of testing in kindergarten and first grade, this was accomplished by using the same sets of assessments in each round, with alternative overlapping secondstage forms. The third- and fifth-grade assessments used the same overlapping two-stage design but with more advanced sets of items. And in eighth grade, the same overlapping two-stage design was used, with two instead of three second-stage forms as were used in prior rounds. Putting K-1, third-, fifth-, and eighth-grade scores on a common scale required common items shared between subsequent assessments. Items from the K-1 assessments (22 in reading and 14 in mathematics) provided the necessary link between K-1 and third grade, with a small "bridge" sample of second-graders augmenting the gap in ability levels between first and third grade. Overlapping ability distributions for third- and fifth-grade made a fourth-grade bridge sample unnecessary. Fifth-grade items shared with the third-grade assessment (59 common items in reading, 31 in mathematics, and 27 in science) supported the extension of the K-1-3 longitudinal scale through fifth grade. Similarly, the eighth-grade items shared with those from fifth grade (17 common items in reading and in science, and 24 in mathematics) extended the longitudinal scale from kindergarten through eighth grade. Eighth-grade tests contained fewer items than the earlier rounds, primarily because the items were, on average, harder and took longer; consequently, there were fewer common items shared between fifth and eighth grades than had been the case previously.

2.1.4.5 Curriculum Relevance

Both eighth- and tenth-graders participated in the 2006 field test of cognitive items. Although there was no tenth-grade round of data collection, the tenth-grade field test data did play a role in the design of the test forms for the eighth grade. Analysis of field test data was carried out for both grades combined, as well as separately for eighth grade and tenth grade. In selecting items for the eighth-grade test forms, in order to avoid ceiling effects, some items selected in the field test for tenth grade were included on the high second-stage form for the eighth-grade assessments. Conversely, items showing a similar percent correct in both eighth and tenth grades suggested that their content was not emphasized in tenth-grade curriculum materials and, therefore, would be appropriate for the eighth-grade forms.

2.1.4.6 Framework Specifications

Items were selected to match the target percentages specified in the framework tables in section 2.1.2 as closely as possible (see tables 2-5 to 2-7). Some compromises in matching target percentages were necessary to satisfy constraints related to other issues, including linking to the earlier rounds, avoiding floor and ceiling effects, and maintaining item quality. This was especially true for the reading assessment in which several questions based on each reading passage placed an additional constraint on the selection of items to match content strands. Reading items were not selected individually but in sets of three to nine items based on the reading passages. Once an investment of time had been made reading a passage, accuracy of measurement per unit of time could be maximized by selecting as many high-quality items as possible based on the passage, even if that resulted in overrepresentation of a content strand. Conversely, a shortfall in a content strand could result if the available items in the strand were linked to a reading passage that had too few other useful items to justify its selection.

Table 2-5.Reading eighth-grade framework targets and percent of assessment items: School year2006–07

				Initial understanding/		
Percent of	T 1	Basic	** 1 1	developing	Personal	Critical
assessment items	Total	Sk1lls	Vocabulary	interpretation	reflection	stance
Target	100	0	0	55	15	30
Actual	100	0	0	73	7	20

NOTE: The framework categories of initial understanding and developing interpretation were combined in the eighth-grade design, but were separated in prior rounds.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2006.

Table 2-6.Mathematics eighth-grade framework targets and percent of assessment items: School year2006–07

Percent of		Number sense,		Geometry and	Data analysis,	Patterns,
items	Total	operations	Measurement	spatial sense	probability	functions
Target	100	20	15	20	15	30
Actual	100	20	15	20	15	30

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2006.

Table 2-7.Science eighth-grade framework targets and percent of assessment items: School year2006–07

Percent of				
assessment				
items	Total	Earth and space science	Physical science	Life science
Target	100	40	30	30
Actual	100	39	31	31

NOTE: Detail may not sum to totals due to rounding.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2006.

While the NAEP frameworks were used as a basis for establishing the content of the ECLS-K assessments at eighth grade, there are distinct differences in the numbers of items in the two assessments. The total testing time available to respond to all cognitive questions in a subject at one grade level in NAEP is longer than that for the ECLS-K. NAEP tests a great many items in the pool, spiraling them among multiple forms, without being constrained as in the ECLS-K. In assembling the questions into forms for the ECLS-K, two important goals were carefully balanced. One was the need to maximize reliability of measurement by administering as many questions as possible to each child within the available time, and the second goal was to include questions that span a range of difficulty, even though the more difficult questions will require more time to answer.

Striking this balance (in addition to satisfying the other goals and constraints in this study) resulted in a difference between the actual and targeted percentages for the content categories in the eighth-grade reading forms for several reasons. First, personal reflection and critical stance content categories typically require additional time for response (see section 2.1.2.1). Conversely, the items from the content categories of initial understanding and developing interpretation may not require as much time for a response, and therefore more items can be included in a shorter period of time, resulting in greater accuracy in estimation of ability. Difficulty in increasing the number of critical stance items also was due to the limited number of critical stance items in most passages. All available passages with multiple critical stance items were selected for the national forms, with the exception of one passage whose items showed poor discrimination in the field test. Of the nine passages selected for the national forms, five contained critical stance items, while the remaining four passages were selected for other purposes (i.e., routing, overlap with fifth grade, difficult items for high form). Similarly, each passage contained only a single personal reflection item. Three of the nine passages selected for the national forms contained personal reflection items, while the remaining six were selected for other purposes (i.e., easy items for the low form, overlap with fifth grade, proficiency levels). Increasing the number of personal reflection and critical stance items would require the addition of other passages to the national forms and result in increased assessment time; therefore, the content percentages were accepted as compromises necessary to satisfy multiple constraints.

Item selections for the mathematics and science assessments closely matched framework target percentages, in large part because the constraint of selecting items in groups was not present. Enough high-quality mathematics items were available for selection in each of the content strands to match frameworks exactly. Minor deviations from framework targets in science are a result primarily of the total number of items administered. The science targets do not match by approximately 1 percent, higher in earth science and lower in physical and life sciences.

The deviations from framework targets probably have relatively little impact on the measurement of the domain of interest because there is some ambiguity in the classification of items. Many if not most of the eighth-grade reading and mathematics items had aspects of multiple skills. For example, answering a reading comprehension item would require decoding the words in the story, understanding the meaning of words in context, and using personal experience to interpret the reading

passage and the question. Similarly, a graph-reading item in the mathematics assessment could be classified as data analysis, statistics, and probability but would also require an understanding of numbers. Therefore, the designation of a single strand category for each item was somewhat arbitrary. It was unlikely that the necessary compromises in selecting items would have had a serious negative impact on measurement of the intended construct.

2.1.4.7 Practical Issues

The 80-minute time allocation for the eighth-grade direct cognitive assessments was divided into 40 minutes for reading and 20 minutes each for mathematics and science. Analysis of field test timings showed that more time per item was needed for reading, due to the extra time required for reading the passages before answering the questions, and less for mathematics and science questions. The sets of mathematics and science items, consisting of short-answer questions, tended to go much more quickly. The number of items in each of the eighth-grade test forms is shown in table 2-8.

Routing test cut points were determined empirically based on field test IRT ability estimates and item parameters. Using the ability estimates for field-tested eighth-graders, and those estimated to be off-grade, simulations were carried out to predict, for each child, a score on the items selected for the routing test and a predicted score on each of the two proposed second-stage forms. Cross-tabulations of the simulated routing scores against each second-stage score were examined, and routing cut points were selected such that ceiling and floor effects would be minimized. This procedure was carried out rather than relying on cut points that approximated a possible 50-50 percent assignment to second-stage forms because it was more important for children to receive test questions matched to their ability than it was to achieve a particular distribution of test forms. Table 2-8 shows the cut scores for each routing test. Sections on samples and operating characteristics in chapter 4 (sections 4.3.1, 4.4.1, and 4.5.1) show the actual percentages achieved in the assessment of the eighth-grade longitudinal sample. The success of the two-stage test design in achieving its goals is discussed there as well.

Description	Reading	Mathematics	Science
Number of items per form			
Routing test	10	10	10
Low second-stage form	19	20	17
High second-stage form	21	20	17
Total item pool	212	174	111
Common items (total)	81	49	38
K-1 and third grade	13	9	ţ
First-grade supplement and fifth grade	2	Ť	†
Third and fifth grade	40	17	21
K-1 (or first-grade supplement), third, and fifth grade	9	4	Ť
Fifth and eighth grade	12	9	11
Third, fifth, and eighth grade	5	10	6
Unique items (total)	131	125	73
K-1 only (including first-grade supplement)	68	50	†
Third grade only	16	34	35
Fifth grade only	21	20	19
Eighth grade only	26	21	19
Routing test cut scores			
Route to low second-stage form	0–5	0–6	0–4
Route to high second-stage form	6-10	7-10	5-10

Table 2-8.Number of items in eighth-grade test forms and routing test cut scores, by domain: School
year 2006–07

† Not applicable.

NOTE: The number of items in each eighth-grade pool is less than the sum of the items in the test forms because there is some overlap of items across forms. Four fifth-grade reading items were calibrated but deleted from the final score scale to align the scale with the framework, and one was deleted from scoring because of differential item functioning (DIF) in the fifth-grade sample. Two reading items that had not been scored in third grade because they proved to be too difficult to provide useful information for third-graders performed satisfactorily when fifth-grade responses were added to the analysis. These two items, present but not scored in third grade, were added to the longitudinal scale. Similarly, one mathematics item that had unsatisfactory statistics in third grade was added to the longitudinal scale based on the combined third-and fifth-grade data. See chapters 4 and 5 for details.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

2.2 Socioemotional Development Measures

In the third-grade and the fifth-grade data collections, children rated their own academic competence and social skills. The self-description questionnaire was designed to determine how children feel about themselves both socially and academically. A literature review on social and emotional development in grades 2 through 5 (Atkins-Burnett and Meisels 2001) indicated the centrality of self-

concept. Examination of different instruments used to assess social and emotional development in grades 2 through 5 led to a recommendation to include several scales from the Self Description Questionnaire I (SDQ I) (Marsh 1992) in the assessment battery (Atkins-Burnett and Meisels 2001). The SDQ I assesses self-concept multidimensionally. Four of the subscales from the SDQ I were included in the spring 2000 and spring 2002 field tests: Reading, Mathematics, All School Subjects, and Peer. The children responded to the SDQ I questions prior to the administration of the cognitive assessment. The response scale as well as several of the items were adapted for use, with permission from the test publishers, in the main study and administered in the third- and fifth-grade data collection periods.

The original SDQ I has some negatively worded items that were not scored, but were included in the instrument so that respondents were not only responding to a set of positively worded items. ECLS-K items asking about problem behaviors were substituted for these items (Atkins-Burnett and Meisels 2001). Problem behavior items served the dual purposes of breaking any response sets and gathering information about the child's perception of behaviors that might interfere with learning. Items measuring both internalizing and externalizing problem behaviors were included. The internalizing problem behavior items assessed acting out behaviors and attention problems.

In eighth grade, a new version of the self-description questionnaire was developed using items from a published instrument designed for adolescents (Self Description Questionnaire II) (Marsh 1992b). The Content Review Panel reviewed the scales available on the SDQ II and determined that Perceived Interest/Competence in Reading and the Perceived Interest/Competence in Math were the most appropriate for the eighth-grade student questionnaire. To capture internalizing behavior problems, the Content Review Panel recommended that the internalizing problem behavior scale used in third and fifth grade be retained for the eighth-grade data collection.

The self-description questionnaire consists of 16 statements. Children rated whether each item was "not at all true," "a little bit true," "mostly true," or "very true." Three subscales were produced from the self-description questionnaire items. The scale scores on all self-description questionnaire scales represent the mean rating of the items included in the scale.

• The SDQ Perceived Interest/Competence–Reading subscale includes four items on grades in English and the child's interest in and enjoyment of reading.

- The SDQ Perceived Interest/Competence–Math subscale includes four items on mathematics grades and the child's interest in and enjoyment of mathematics.
- The SDQ Internalizing Behavior subscale includes eight items on internalizing problem behaviors such as feeling "sad a lot of the time," feeling lonely, feeling ashamed of mistakes, feeling frustrated, and worrying about school and friendships.

To measure the children's self concept, the Content Review Panel recommended a second set of scales for the eighth-grade student questionnaire. The Self-Concept and Locus of Control scales were adapted from the National Education Longitudinal Study of 1988 (NELS:88). The Self-Concept scale comes from the Rosenberg Self-Esteem Scale (RSE) (Rosenberg 1965). These scales asked children about their perceptions about themselves and the amount of control they had of their own lives. Items were drawn from the NELS:88 student questionnaire and asked children to indicate the degree to which they agreed with 13 statements about themselves. Items from the Self-Concept scale included "I feel good about myself," "I feel I am a person of worth, the equal of other people," and "I feel I do not have much to be proud of." Items from the Locus of Control scale included "I don't have enough control over the direction my life is taking" and "When I make plans, I am almost certain I can make them work." They chose from the following responses: "strongly agree," "disagree," or "strongly disagree" for each item.

3. ANALYSIS METHODOLOGY

This chapter describes the procedures used in processing the ECLS-K eighth-grade assessment data and producing scores for analysis and for inclusion in user files. Quality control steps are described in section 3.1, followed by an explanation of the methodology used to carry out specialized procedures for psychometric analysis. A three-parameter item response theory (IRT) model was used to put scores obtained on different assessment forms on the same scale for the purpose of comparisons within and across assessment years. Differential item functioning (DIF) procedures identified test items that performed differently for subgroups of the population. The development of longitudinal score scales is described in chapter 5.

3.1 Quality Control Procedures

Procedures employed to ensure accuracy in the collection of the cognitive test item data are described in section 4.6 of the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99* (ECLS-K), Eighth-Grade Methodology Report (NCES 2009–003) (Tourangeau et al. forthcoming). In the subsequent steps of converting the resulting raw item response data to final scores, procedures were checked to ensure the accuracy and validity of the results. A series of steps were carried out, from converting raw examinee item responses into scores for individual items, to evaluating item functioning using both classical item analysis and IRT methods, to assembling item data into meaningful and interpretable scores. Throughout the process, attention was given both to checking that steps were carried out correctly, and to verifying that results accurately represented the constructs they were designed to measure. The procedures described and utilized represent the standard analysis procedures used on all longitudinal studies of this type.

Frequency distributions of raw examinee item responses were produced for each test item to serve as a baseline for confirming the accuracy of later processing steps. Each distribution was compared with the text of the corresponding question in the assessment, and with the scoring instructions, to confirm that responses were coded as expected. For example, for a four-option multiple choice question, the data file would be expected to contain response codes of 1, 2, 3, and 4, while 1 (correct) or 2 (incorrect) was to have been recorded by the scorer for open-ended questions. A missing data code for omitted or unreadable responses was also counted for each item.

Within each subject area, children who had not responded to enough test items to receive a score were identified. "Too few items" was defined as answering fewer than 10 questions in the routing and second-stage forms combined. Only items actually attempted by the child were counted toward the scorability threshold. Before being deleted from further analysis, each "too few items" data record was reviewed visually to verify that not enough valid item responses were present. On the reading assessment, 82 children responded to fewer than 10 items, with 22 children on the mathematics assessment, and three children on the science assessment.

Classical item analysis was carried out for each test form (routing test and second-stage forms separately) using Educational Testing Service's (ETS) proprietary software, F4STAT. Sets of statistics were produced for each item, as well as summary statistics for the section as a whole. Each of these statistics provides information on item performance, as well as a source of quality control data. For each item, the number and percentage of test takers choosing each response option is computed, as well as their average number of correct answers on the whole test section. The same statistics are computed for children who omitted the item (and answered at least one subsequent item) and for those who did not answer the item *or* any subsequent items ("not-reached"). The response frequencies from the item analysis procedure were checked, item by item, against the baseline response frequencies initially obtained on the raw data file to confirm that responses and missing data codes had been interpreted correctly.

Summary statistics for each item include P+ (percent correct) and *r*-biserial (the correlation of item score with total test score, adjusted for the item score being dichotomous). These statistics were reviewed to verify that an unambiguous correct answer key was used for each item, meaning not only that the *intended* right answer was tagged, but that the tagged answer was in fact functioning as an unambiguous right answer. Evidence for the validity of the answer key comes from two sources: the mean average section score for test takers choosing the correct response should be higher than that of the groups choosing incorrect responses; and the *r*-biserial should be positive, ideally at least 0.30 or higher. If these conditions are not satisfied, one of two error conditions could be responsible. An incorrect answer key could have inadvertently been applied 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. Because all of the eighth-grade items had been used in previous assessments or field tested, and the response options had been evaluated, no flawed items were found.

Items within each test section had been arranged in ascending order of anticipated difficulty. A review of the item P+s would identify any serious deviation from this expectation, which could indicate anomalies in the administration or scoring of items. Similarly, unexpectedly large omit or not-reached counts for an item or items could call into question whether routing steps were applied correctly. No such indicators of data or administration errors were detected in reviewing item analysis tables.

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

Frequency distributions of routing test scores were compared with the distributions for each second stage form to confirm that the routing had been carried out at the correct cut points, i.e., that the number of observations for each second-stage form matched the number in the corresponding score range of the routing test. Data records were reviewed visually to confirm that the discrepant counts (e.g., number routed to the low form vs. the number who answered one or more items on the low form) reflected what was actually in the raw data files. The change from computer-assisted administration to paper-and-pencil tests, with routing sections hand-scored on-site, introduced another possible source of error in the eighth grade. Unlike the earlier rounds, in which routing test responses were scored by the computer and second-stage forms automatically selected, the eighth-grade routing tests were scored, and second-stage forms selected, by the test administrators. For a very small number of children (73 in reading, 48 in mathematics, and 60 in science), an incorrect second-stage form was chosen, and in a few cases, second-stage item responses were present in the data file although routing items were not (15 in reading, 36 in mathematics, and 37 in science). In these cases, it was not possible to determine whether the routing test had been administered but responses had not made it onto the data file, or a second-stage form had been selected without a routing test, possibly due to time constraints. (Review of the standard errors of the estimates showed that the scores estimated for these children were not negatively impacted by these anomalies.) All data records were scored, as long as they had at least the minimum of 10 item responses on either or both forms.

Frequency distributions of total number correct (routing plus second stage combined) were examined separately for each form combination (i.e., routing+low form, routing+high form) to look for possible floor and ceiling effects. While this is not a quality control issue in the sense of verifying the accuracy of the scoring procedures, it does have implications for interpretation and analysis of the resulting scores. A floor effect occurs when the test is too difficult overall for some test takers, and the score distribution contains a substantial number of children scoring at the chance, or guessing, level. Conversely, a test with a ceiling effect is too easy for some children and a substantial number are able to answer all, or nearly all, of the items correctly. Slight floor effects in the eighth grade tests were observed, and are discussed in chapter 4.

The next step in processing the raw item responses was preparing scored item files for input to the IRT calibration procedures, that is, replacing raw response option codes (e.g., 1, 2, 3, 4) with standard codes for correct, incorrect, omitted, and not reached items (1, 0, 2, and 3, respectively). Omitted items were defined as unanswered items that were followed by a response to at least one subsequent item, while unanswered items coded as "not reached" had no subsequent items answered. The quality control procedure for confirming that this was done correctly consisted of printing, for a spaced sample of every 100th case, the raw and scored data record, along with the answer keys, and hand-checking the conversions. In some cases, additional records were needed, so that all variations found in the raw data file could be checked. For example, if the spaced sample of quality control records happened to have only cases that were routed to the low second-stage form, additional records were obtained so that high form score conversions could be verified as well. Producing the scored item files entailed reorganizing the order of test items, because some items appeared in more than one second-stage form. In order to strengthen the linkage of each set of forms to the same scale, the scores for these common items needed to be relocated from their original separate locations to a single common location. An item map was developed to direct the reordering of the common items. Scores that were simple sums of number correct on a specified set of items (e.g., reading and mathematics proficiency level scores: see section 4.1.4 for definitions) were computed at this time, checked for the same spaced sample, and inserted into the scored item records. Although number-right proficiency scores do not appear in the user files because the sets of items were not taken by all test takers, the number-right counts for the proficiency levels were needed as input to the IRT calibration step. The eighth-grade scored item files were then combined with the scored item files from kindergarten through fifth grade. Like the test items shared in common across test forms within eighth grade, items shared in common across rounds were positioned together for IRT calibration, and again, frequency counts were checked to confirm the accuracy of the files.

Finally, item-by-item frequency distributions were produced for the scored, reordered files; for the common items, the frequency counts were checked against the aggregates of the frequencies for the separate forms and rounds in which the items originally appeared. These frequency counts, and item means computed on the verified scored item file, provided the basis for checking the results of the IRT scaling steps.

Section 3.2 below describes PARSCALE, the IRT program used for calibrating item parameters and test takers' ability levels on a scale that is then used to produce scale scores on the whole item pool and probability scores for the proficiency levels. Statistics and graphs produced by the PARSCALE program and its associated graphing program (PARPLOT) were used not only to verify the accuracy of the computations, but also to evaluate the reasonableness of the results.

PARSCALE produces counts, for each test item, of the number of responses, number of omits, number right, and number wrong found in the input scored data file. The percent correct for each item is also computed. These counts and percents 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 interpreted correctly.

Another perspective on quality assurance, aside from verifying the accuracy of data and computations, is the extent to which the scoring model appropriately represents the information in the whole item pool. The *r*-biserials produced in the classical item analysis steps show the relationship of each test item to the rest of the form on which it appears. The IRT "a" parameter, and the PARSCALE plots, demonstrate the cohesiveness of the *whole set* of items used in kindergarten through eighth grade in each subject (or for science, third to eighth grade only). High "a" parameters (1.0 or above) mean that items were strongly related to the underlying construct represented by the item pool. Nearly all reading and mathematics items had "a" parameters, as would be expected for a pool of items that are less strongly related to each other.

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 for grouped points across the range of estimated ability levels are superimposed on the same graph. The closeness of fit of the data to the logistic function can be interpreted as confirming the appropriateness of the IRT model for scoring the tests. More detail on the IRT model is presented in section 3.2, and a full description of the use and evaluation of the IRT procedures in developing the longitudinal scale appears in chapter 5.

The final steps in producing the IRT-based scores consisted of aggregating probabilities of correct responses across the whole item pool in each subject for the scale scores and obtaining weighted means of ability estimates for standardized scores that represented population estimates at each round. These 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 checking step, means and standard deviations of the final score record were obtained and found to be consistent with expectations. For the scale scores, that would be scale score means that increased from round to round, with ranges that were consistent with the number of items in the pool for each subject. The standardized scores could be explicitly checked, since by definition their weighted mean should equal 50.0 with a standard deviation 10.0 within each round.

3.2 Overview: The Three-Parameter Model

Measuring the extent of cognitive gains at both the group and individual level requires that the various kindergarten through eighth-grade assessment forms be calibrated on the same scale. The most convenient way of doing this is to use IRT. To successfully carry out such a calibration, the sets of test items should be relatively unifactorial within a subject area (reading, mathematics, or science), with the same dominant factor underlying all test forms. This suggests that there should be a common set of items across adjacent forms and that most, but not necessarily all, content strands be represented in all grade forms. Increments in difficulty demanded in ascending grade forms (kindergarten through eighth grade) can be accomplished by (1) increasing the problem-solving demands within the same content areas and (2) including content in the later forms (in particular fifth and eighth grade) that measures materials normally found in the curriculum for higher grades and that builds on skills learned in earlier grades.

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

3.2.1 Overview of Item Response Theory

The underlying assumption of IRT is that a test taker's probability of answering an item correctly is a function of his or her ability level for the construct being measured and of one or more characteristics of the test item itself. The three-parameter IRT logistic model uses the pattern of right, wrong, and omitted responses to the items administered in a test form and the difficulty, discrimination power, and probability of guessing correctly, given the lowest level of ability, of each item, to place each test taker at a particular point, θ (theta), on a continuous ability scale. Figure 3-1 is an example of a graph of the logistic function for a hypothetical test item. The horizontal axis represents the ability scale, theta. Points along the vertical axis represent the probabilities of answering an 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* as

$$P_{i}(\theta) = c_{i} + \frac{(1 - c_{i})}{1 + e^{-1.702^{*}a_{i}(\theta - b_{i})}},$$
(3.1)

where θ = ability of the test taker;

 $a_i =$ discrimination of item i, or how well changes in ability level predict changes in the probability of answering the item correctly, at a particular point;

 $b_i = difficulty of item i; and$

 c_i = "guessability" of item i, that is, the probability that a very low-ability test taker will answer item i correctly.

The "c" parameter represents the probability that a test taker with very low ability will answer the item correctly and is generally a function of the number of available response options. In figure 3-1, about 20 percent of test takers with a very low level of mastery of the test material guessed the correct answer to the question. The c parameter will not necessarily be equal to 1/(number of options) (e.g., .25 for a four-choice item). Some response options may, for unknown reasons, be more attractive than random guessing, while others may be less likely to be chosen.

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



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

The IRT "b" parameters correspond to the difficulty of the items, represented by the horizontal axis in the ability metric. In figure 3-1, b = 0.0 means that test takers with an estimated ability $\theta = 0.0$ have a probability of getting the answer correct that is equal to halfway between the guessing parameter and 1. In this example, 60 percent of people at this ability level would be expected to answer the question correctly. The "b" parameter also corresponds to the point of inflection of the logistic function. This point occurs farther to the right for more difficult items and farther to the left for easier ones. Figure 3-2 is an example of a graph of the logistic functions for seven different test items, all with the same "a" and "c" parameters and with difficulties ranging from b = -1.5 to b = 1.5. For each of these hypothetical questions, 60 percent of test takers whose ability level matches the difficulty of the item are likely to answer correctly. Fewer than 60 percent will answer correctly at values of theta (ability) that are less than "b," and more than 60 percent at $\theta > b$.

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



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

The discrimination parameter, "a," has perhaps the least intuitive interpretation of the three IRT parameters. It is proportional to the slope of the logistic function at the point of inflection. Items with a very steep slope are said to discriminate well. In other words, they do a good job of discriminating, or separating, people whose ability level is below the calibrated difficulty of the item (who are much less likely to get it right) from those of ability higher than the item "b," who are much more likely to answer correctly. By contrast, an item with a relatively flat slope is of little use in determining whether a person's correct placement along the continuum of ability is above or below the difficulty of the item. This idea is illustrated by figure 3-3, representing the logistic functions for two test items having the same difficulty and guessing parameters but different discrimination. The test item with the steeper slope (a = 2.0)provides useful information with respect to whether a particular test taker's ability level is above or below the difficulty level, 1.0, of the item: if the answer to this item was incorrect, the person very likely has an ability below 1.0; if the answer was correct, the test taker probably has a θ greater than 1.0, or guessed successfully. A series of many such highly discriminating items, with a range of difficulty levels (b parameters) such as those shown in figure 3-2, will do a good job in narrowing the choice of probable ability level. Conversely, the flatter curve in figure 3-3 represents a test item with a low discrimination parameter (a = 0.3). There is little difference in proportion of correct answers for test takers several points apart on the range of ability. In this example, knowing whether a person's response to such an item is correct or not contributes relatively little to pinpointing his or her correct location on the horizontal ability axis.

With respect to evaluating item quality, "a" parameters (the discrimination parameter) should each be over 0.50. Items with "a" parameters of 1.0 or above are considered very good. As described earlier, the "a" parameter indicates the usefulness of the item in discriminating between points on the ability scale. The "b" parameters, or item difficulties for the items, should span the range of abilities being measured. Item difficulties should be concentrated in the range of abilities that contains most of the test takers. Test items provide the most information when their difficulty is close to the ability level of the examinees. Items that are too easy or too difficult for most of the test takers are of little use in discriminating among them. Ideally, the "c" parameters (the probability of a low ability person guessing correctly) tend to be about .25 or less for four-choice items, but they may vary with difficulty and, of course, the number of options. Open-ended items typically have a "c" parameter that is close to 0. In general, the ECLS-K item parameters met these standards.

Figure 3-3. Three-parameter IRT logistic functions for two hypothetical test items with different discrimination (a)



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

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

In addition to providing a mechanism for estimating scores on items that were not administered to every individual, IRT has advantages over raw number-right scoring in the treatment of guessed and omitted items. By using the overall pattern of right and wrong responses to estimate ability, the model gives very little credit for correct answers to hard items by low ability children. Omitted items are treated as if the examinee had guessed at random.

3.2.2 Item Response Theory Estimation Using PARSCALE

The PARSCALE (Muraki and Bock 1991) computer program computes marginal maximumlikelihood estimates of IRT parameters that best fit the responses given by the test takers. The procedure estimates "a," "b," and "c" parameters for each test item, iterating until convergence when a specified level of accuracy is reached. Comparison of the IRT-estimated probability of a correct response with the actual proportion of correct answers to a test item for examinees grouped by ability provides a means of evaluating the appropriateness of the model for the set of test data for which it is being used. A close match between the IRT-estimated probabilities and the empirical proportion correct means that the theoretical model accurately represents the empirical data.

As indicated earlier, a longitudinal growth study by its very nature consists of subpopulations defined by differing ability levels. That is, after all the kindergarten, first-grade, third-grade, fifth-grade, and eighth-grade assessments had been completed (seven rounds, counting fall and spring administrations in kindergarten and first grade) there were seven recognizable subpopulations of different ability levels, which are tied to the time of testing. For example, the fall-kindergarten subpopulation will have, on average, a lower expected level of performance than that found in each of the remaining followups. Similarly, the average performance of the fall-first graders will be lower than that of the same children the following spring. The bridge sample of second-graders, designed to fill in the gap in testing between first and third grade, represents an eighth subpopulation.

When the first round of kindergarten data was collected in fall 1998, relatively few children were routed to the middle-level second-stage forms and even fewer to the high-level forms. Thus, there were not enough observations on the most difficult items to obtain stable item parameter estimates. As the children were retested in spring-kindergarten and fall- and spring-first grade the following year, more and more data could be used to stabilize the estimates for the middle- and then the high-level items. The same is true for the most difficult first-grade items that were repeated in third grade, for third-grade items repeated in fifth grade, and for fifth-grade items repeated in eighth grade. As each round of data became available, item responses were pooled and parameters re-estimated. The pooling of all time points and reestimating the item parameters, of course, results in a remaking of history in a longitudinal study where intermediate results are published before all the data from all the time periods are available. That is, falland spring-kindergarten scores that have been reported and analyzed were later modified somewhat when first-grade data became available. Similarly, all kindergarten and first-grade scores were replaced when the scale was extended to incorporate the third-grade assessment items; all kindergarten, first-grade, and third-grade scores were replaced to include fifth-grade data; and now, with the addition of eighth-grade items to the scales, scores from all previous rounds were re-estimated. The use of all data points over time is desirable because it can provide updated estimates of both the item and latent ability parameters throughout the entire ability distribution on a vertical scale. This procedure was used in the vertical scaling that was carried out for the National Education Longitudinal Study (NELS:88) (Rock et al. 1995) and for High School and Beyond (Rock et al. 1985; Rock and Pollack 1987).

A strength of the PARSCALE and other Bayesian approaches to IRT is that they can incorporate prior information about the ability distribution (i.e., from the round of data collection from which an observation is taken) in the ability estimates. This is particularly crucial for measuring change in longitudinal studies. It provides an acceptable way of coping with perfect and chance scores (i.e., correct answers to all items administered or scores at the guessing level or below, respectively). For example, a few very advanced individuals who took the high-level mathematics form in spring-first grade might get all the items correct. These individuals, while gifted, may not get perfect scores when they eventually are tested on a harder set of items in later grades. Will this mean that they are less skilled in later grades than in first grade? Probably not. Conversely, individuals scoring at or below the chance level at two time periods may have gained skills that are below the level assessed by the test items. Pooling all available information, that is, pooling all item responses for all people at all time points, and recalibrating all of the item parameters using Bayesian priors (updated in light of new observations) reflecting the ability distributions associated with each particular round, provides for an empirically based narrowing of the

distribution toward the mean such that the extreme item parameters "shrink" to more reasonable item parameters and ability scores (Muraki and Bock 1991).

The fact that the total item pool is used in conjunction with the Bayesian priors leads to shrinking back the extreme item parameters, as well as the perfect and chance scores, which in turn allows for the potential of some gains even in the upper and lower tails of the distribution. Each of the rounds of data collection in kindergarten through eighth grade is treated as a separate subpopulation with its own ability distribution. The amount of shrinkage in ability estimates is a function of its distance from the subgroup mean distributions 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 those that are near the mean). Theoretically this approach has much to recommend it. In practice, it has to have reasonable estimates of the difference in ability levels among the subpopulations in order to incorporate realistic priors. Essentially, the scales are determined by the linking items (i.e., the items common to the item batteries used in different rounds of data collection) and the initial prior means for the subgroups are in turn determined by the differential performance of the subpopulations on these linking items. For this reason the item pool has been designed to have an overabundance of items linking the forms. This approach, using adaptive testing procedures combined with Bayesian procedures that allow for priors on both ability distributions and on the item parameters, is needed in longitudinal studies to minimize floor and ceiling effects.

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

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

$$L(\beta) = \prod_{g} \prod_{j:g} \int_{\theta} P(x_{j:g} | \theta, \beta) f_{g}(\theta) d(\theta)$$

$$\approx \prod_{g} \prod_{j:g} \sum_{k} P(x_{j:g} | \theta = X_{k}, \beta) A_{g}(X_{k}).$$
(3.2)

In equation (3.2), $P(x_{j:g}|\theta,\beta)$ is the conditional probability of observing a response vector $x_{j:g}$ of person j from group g, given proficiency θ and vector of item parameters $\beta = (a_1, b_1, c_1, ..., a_k, b_k, c_k)$, and $f_g(\theta)$ is a population density for θ in group g. Prior distributions on item parameters can be specified and used to obtain Bayes modal estimates of these parameters (Mislevy 1984). The proficiency densities $(d(\theta))$ can be assumed known and held fixed during item parameter estimation or can be estimated concurrently with item parameters.

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

Maximization of $L(\beta)$ is carried out by an application of an EM algorithm (Dempster, Laird, and Rubin 1977). When population densities are assumed known and held constant during estimation, the algorithm proceeds as follows. In the E step, provisional estimates of item parameters and the assumed multinomial probabilities are used to estimate expected sample sizes at each quadrature point for each group (denoted \hat{N}_{gk}), as well as over all groups (denoted $\hat{N}_k = \sum_g \hat{N}_{gk}$). 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 step, improved estimates of the item parameters, β , are obtained using maximum likelihood by treating the \hat{N}_{gk} and \hat{r}_{ik} as known, subject to any constraints associated with prior distributions specified for β . The user of the multiple group version of PARSCALE has the option of fixing the priors on the ability distribution or allowing the posterior estimate to update the previous prior and combine with the data-based likelihood to arrive at a new set of posterior estimates after each major EM cycle. If one wishes to update on each cycle, one can continue to constrain the priors to be normal or their shape can be allowed to vary. The ECLS-K approach was to allow for updating the prior but with the normality assumption. The smoothing that came from the updated normal priors led to less jagged-looking ability distributions and did not tend to overfit the item parameters. Lack of fit in the item parameter distribution would simply be absorbed in the shape of the ability distribution if the updated ability distribution were allowed to take any shape. A similar procedure was used in estimating the item parameters in the National Adult Literacy Study (NALS) (Kirsch et al. 1993).

It should be remembered that the solution to equation 3.2 finds those item parameters that maximize the likelihood across all eight time points (the seven longitudinal ECLS-K rounds plus the second-grade bridge sample). The present version of the multiple group PARSCALE saves only the subpopulation means and standard deviations and not the individual expected *a posteriori* (EAP) scores. The individual EAP scores, which are the means of the posterior distributions of theta, 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 being equal as in the standard conditioning procedure.

Conditional independence is an assumption of all IRT models, but as Mislevy et al. (1992) point out, it is a strong assumption that is often violated in practice. However, if one thinks of IRT-based scores as a summarization of essentially the largest latent factor underlying a given item pool, then small violations are of little significance. To ensure that there were no substantive violations of this assumption, all graphs were inspected to ensure a good fit throughout the ability range. For each item, the empirical proportion correct in each round was computed and compared with the model-based estimated proportion correct based on thetas for the same set of children, that is, the subset of children in the round who had received and responded to the item. Discrepancies between predicted and actual item proportion correct were reviewed for each round. No systematic over- or under-prediction was found for any round or for any type of item.

Tables B1 to B3 in appendix B list the IRT item parameters for the three subject areas. The items are sorted in ascending order of difficulty (the IRT "b" parameter). These tables also show the

assessment versions in which the items appeared: one set of tests used for the first four rounds, fall- and spring-kindergarten and fall- and spring-first grade, with new versions used in third, fifth, and eighth grades. Items that appeared in more than one assessment version served to link the scales across rounds (see section 5.1). Appendix B also shows the mean and standard deviation of the IRT ability estimate, theta, within each round. Bands marking two standard deviations below and above the theta mean illustrate the match of assessment difficulty to the range of child ability in each round.

Tables C1 to C3 in appendix C show estimates of the proportion of correct responses to each item that would have been expected if all children had answered all of the items in the kindergarten through eighth-grade item pools at every round. Although each child answered only a small subset of the items each time, IRT ability estimates and item parameters make it possible to estimate performance on all of the items in the pool. In appendix D, tables D1 to D3 show the fit of the IRT model to the item response data. The IRT-estimated probability of a correct response was calculated for each item answered by each child. The average of these probabilities is equivalent to the estimated proportion correct predicted by the IRT model for each answered item. These estimates were compared with the actual proportion correct observed for the answered items. The tables in appendix D show the differences for each item (actual minus predicted), for all items used in each round. In addition to comparisons of predicted and actual proportion correct, the IRT logistic function graphs for each round are compared for common functionality. This inspection includes review of discrepancies in the item fit to the model across rounds for all ability ranges; determination of a systematic over- or under-prediction of the model; unusual features in the data at the extremes of the distribution; and overall comparability of the data across rounds. For nearly all items in nearly all rounds, these discrepancies were small, indicating good fit of the IRT model to the item response data.

3.2.3 Standard Errors of Measurement Using the Information Function

In statistics and psychometrics, the precision of parameter estimates can be measured using the information function. This is computed as a function of the reciprocal of the measurement error, or the variability of repeated estimates of the value of the parameter, denoted as σ^2 . Thus, the less measurement

error is present, the more precise the estimate of the value of a parameter, and the greater the value of the information function. Equation 3.3 defines the information function (I):

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

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

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

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

where $I(\theta) =$ amount of test information at child's ability level θ ; $I_i(\theta) =$ amount of test information at child's ability level θ for item i; and n = number of items answered by the child.

The test information function will be much greater than any single item information function; thus a test measures ability more precisely than does a single item. The test information function is calculated using only the administered items with valid responses. The more items *answered*, then the greater the precision in estimating the ability.
The definition of the item information function depends upon the IRT model used. For the three parameter (a, b, and c) model used in the ECLS-K estimates and described above, the item information function is defined as

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

where $P_i(\theta) = c + (1-c) \frac{1}{1+e^{-L}};$ $L = a(\theta-b); and$ $Q_i(\theta) = 1.0 - P_i(\theta).$

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

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

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

The procedure above was carried out to calculate a standard error for each of the theta estimates in the eighth-grade round of analysis. These standard errors are reported in the data files for each of the thetas in reading, mathematics, and science. The results of the standard error calculations from the eighth-grade analysis for all rounds are presented in chapter 4.

3.3 Differential Item Functioning

Differential item functioning (DIF) as defined here attempts to identify those items showing an unexpectedly large difference in item performance between a focal group (e.g., Black children) and a reference group (e.g., White children) when the two groups are "blocked" or matched on their total score. In other words, DIF is conducted to investigate potential bias in an item. It should be noted that any such strictly internal analysis (i.e., without an external criterion) cannot detect bias when that bias pervades all items in the test (Cole and Moss 1989). It can only detect differences in the relationships among items that are anomalous in some group in relation to other groups. In addition, such approaches can only identify the items where there is unexpected differential performance; they cannot directly imply bias. A determination of bias implies not only that differential performance on the item is related to subgroup membership but also that the difference is unfairly associated with subgroup membership. That is, the difference is due to an attribute not related to the construct being measured. As Cole and Moss (1989) point out, items so identified must still be interpreted in light of the intended meaning of the test scores before any conclusion of bias can be drawn. It is not entirely clear how the term "item bias" applies to academic achievement measures given to children with different patterns of exposure to content areas. For example, some children may be in schools where the fifth- through eighth-grade science curriculum emphasizes life science units, while others may have greater exposure to physical science topics. Both groups may have similar total scores in science, but for one group the life science items may be differentially more difficult while the reverse is true for the other group. It is Educational Testing Service's practice to carry out DIF analysis on all tests it designs in order to detect test items with differential performance for subgroups defined by gender and race/ethnicity.

Two DIF methods were used in detecting differential performance of subgroups on the ECLS-K direct cognitive assessments. 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). The two methods complement one another in detecting differential performance. The methods and advantages of using both procedures are discussed below.

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.

Because of the two-stage, multiform design of the ECLS-K assessments, children were assessed with different sets of items, so number-right scores are not based on items of comparable difficulty. Instead, the IRT ability estimate theta was used as the stratifying variable, divided into 41 equally spaced intervals. Accordingly, 41 two-way tables were produced, one for each theta interval. The first dimension of each of the 41 two-way tables is population subgroups (e.g., White children vs. Black children), and the second dimension is passing versus failing on a given item. Thus, the question that the M-H procedure addresses is whether members of the reference group (e.g., White children) who have the same total ability estimate as members of the focal group (e.g., Black children) have the same likelihood of passing the item in question. Although the M-H statistic looks at passing rates for two groups while controlling for total score, no assumption need be made about the shape of the total score distribution for either group. In this case, the chi-square statistic associated with the M-H procedure tests whether the average odds ratio for a test item, aggregated across all 41 score levels, differs from unity (i.e., equal likelihood of passing the item, given the same overall test score).

The M-H procedure provides a statistical test of whether the average odds ratio significantly departs from unity for each item. If the probability that the odds ratio differs from unity due to chance is less than 0.05, then one could say that there is statistical evidence for M-H DIF on the item in question. The problem with this interpretation is twofold. First, a very large number of statistical tests are being performed, one for each item for each pair of subgroups; therefore, low probabilities will be found occasionally even if no DIF is present. Second, if two relatively large samples are involved, statistical significance will be virtually guaranteed.

Given these reservations, ETS has developed an "effect size" estimate that is not sample-size dependent (Dorans and Kulick 2006). Associated with the effect sizes is a letter code that ranges from "A" to "C." It is ETS's experience that effect sizes of 1.5 and higher have practical significance. Effect sizes of this magnitude that are statistically significant are labeled with a "C" (i.e., C-DIF or C-level DIF). Items labeled "A" or "B" do not show statistically significant differential functioning for the two groups being compared or have differences that are too small to be important, respectively.

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

P-DIF uses a weighting function supplied by the standardization, or focal, group to average differences across levels of the matching variable, or ability measurement, theta. The standardization group supplies specific weighting factors at each score level to weight differences in item performance between the focal and reference groups. The item scores used in the standardization methodology algorithm are represented as proportion correct at each score level. 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, as is done in the M-H procedure.

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

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

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

DIF procedures were carried out for the eighth-grade assessment items for six sets of contrast groups: males (reference group) compared with females (focal group), and White children (reference group) compared with four other racial/ethnic groups: Black, Hispanic, Asian, and "Other." There were too few Native American and multiracial children for DIF statistics to be evaluated separately for these groups, and these children were thus combined in the "Other" subgroup. Statistics were computed for each item for which the minimum number of required responses, 300 observations for the smaller group, was available. The results of DIF analysis for the eighth-grade assessment are discussed in chapter 4.

4. PSYCHOMETRIC CHARACTERISTICS OF THE ECLS-K DIRECT COGNITIVE BATTERY

This chapter documents the direct cognitive test results for the eighth-grade round of testing. The types of scores derived from each of the assessments are described, along with the psychometric characteristics of each. (Notes on the development of longitudinal scales appear in chapter 5, along with a discussion of the analysis of gain scores.) Results for the six kindergarten through fifth-grade rounds are reviewed, to the extent that they are relevant to interpretation of eighth-grade results or to the measurement of gain. The numbers of observations in some of the tables in this chapter may differ slightly from the sample totals in the ECLS-K data file. These analyses were carried out prior to final determination of cases eligible for the data file, and a few cases may have been deleted from the files. The psychometric results presented here may also differ from statistics reported in the user's manual. National estimates in this chapter are based on all children who had been tested at each round, using the corresponding cross-sectional weights (C1CW0-C7CW0). Tables in the user's manual are based on the panel sample, that is, the subset of children who participated in all seven rounds of data collection, and the longitudinal panel weight (C1 7SC0). The emphasis in this chapter is on the psychometric characteristics of the tests at each round, while the user's manual is designed to provide a reference for comparison with statistics obtained from secondary analyses, which may typically employ multiple rounds of data. Weighted score statistics for all direct cognitive scores are presented in appendix A, with breakdowns by sex, race/ethnicity, socioeconomic status, and school type.

Intercorrelations among the subject areas, within and across rounds, are presented in the chapter 5 sections on longitudinal measurement and evaluation of the score scales.

4.1 Types of Scores

The scores used to describe children's performance on the direct cognitive assessment include broad-based measures that report performance in each domain as a whole, as well as targeted scores reflecting knowledge of selected content or mastery within a set of hierarchical skill levels. Some of the scores are simple counts of correct answers, while others are based on item response theory (IRT), which uses patterns of correct and incorrect answers to obtain estimates on a vertical scale that may be compared in different assessment forms. Proficiency scores employ both direct counts and IRT-based

methods. The different types of scores that can be used to describe a child's performance on the direct cognitive assessment are described in detail in this chapter. Number-right scores and IRT scale scores measure a child's performance on sets of questions with a broad range of difficulty. Standardized scores (T-scores) report children's performance relative to their peers. Criterion-referenced proficiency scores and item cluster scores evaluate a child's performance with respect to subsets of items that mark specific skills.

4.1.1 Number-Right Scores

Number-right scores are counts of the raw number of items a child answered correctly. These scores are useful for descriptive purposes only for assessments that are the same for all children. However, when these scores are for assessments that differ in difficulty, they are not comparable to each other. For example, a child who took the high-difficulty mathematics second-stage form would probably have gotten more questions correct if he or she had taken the easier low form. For this reason, raw number-right scores are reported only for the first-stage (routing) sections of the assessments, which were the same for all children being assessed using a particular set of instruments (i.e., the kindergarten-first grade (K-1), third-grade, fifth-grade, or eighth-grade version). The routing test in each subject area consisted of sets of items spanning a variety of skills. For example, the reading routing test used for the four kindergarten and first-grade rounds emphasized prereading skills; in third and fifth grades the routing tests contained easy and difficult decoding words, understanding of words in context, and a series of questions based on a reading passage; and in eighth grade the routing test contained only questions based on reading passages. An analyst might use the routing test number-right scores to report actual performance on these particular sets of tasks. Because the same routing test was used for the fallkindergarten through spring-first grade data collections, rounds 1 through 4, score comparisons may be made among these rounds. However, scores on the third-, fifth-, and eighth-grade routing tests were each based on different and more difficult sets of items. The third-, fifth-, and eighth-grade routing test number-right scores should *not* be compared with the kindergarten or first-grade routing test number-right scores, nor with each other.

4.1.2 Item Response Theory Scale Scores; Standardized Scores (T-Scores)

Broad-based scores based on the full set of assessment items in reading, mathematics, and science were calculated using IRT procedures. The IRT scale scores estimate a child's performance on the whole set of assessment questions in each content domain, while standardized scores (T-scores) report children's performance relative to their peers. IRT made it possible to calculate scores that could be compared regardless of which second-stage form a child received. The IRT scale scores reported here represent estimates of the number of items children would have answered correctly at each point in time if they had taken all of the 212 scored questions in all of the first- and second-stage reading forms administered in all rounds, the 174 scored questions in all of the mathematics forms from all rounds, and the 111 third-, fifth-, and eighth-grade science items. (A small number of additional items were administered but not included in scale scores for reasons explained in sections 4.3 and 4.4.) These scores are not integers because they are probabilities of correct answers, summed over all items in the pools. (Scores for different subject areas are not comparable to each other because they are based on different numbers of questions, as well as content that is not necessarily equivalent in difficulty. That is, it would not be correct to assume that a child is doing better in reading than in mathematics because his or her IRT scale score is higher for reading than for mathematics.) A description of IRT methodology may be found in chapter 3. Chapter 5 contains a discussion of the application of IRT to creating longitudinal scores for ECLS-K.

Standardized scores (T-scores) provide norm-referenced measurements of achievement, that is, cross-sectional estimates of achievement *relative to the population as a whole*. A high mean T-score for a particular subgroup indicates that the group's performance is high in comparison with other groups. It does not represent mastery of a particular set of skills, only that the subgroup's mastery level is greater than a comparison group. Similarly, a change in mean T-scores over time reflects a change in the group's status with respect to other groups. In other words, T-scores provide information on *status compared with a child's peers*, while the IRT scale scores and proficiency scores represent *status with respect to achievement on a particular criterion set of assessment items*. The T-scores may be used as an indicator of the extent to which an individual or a subgroup ranks higher or lower than the national average and how much this relative ranking changes over time.

The standardized scores reported in the database are transformations of the IRT theta (ability) estimates, rescaled to a mean of 50 and standard deviation of 10 using cross-sectional sample weights for each wave of data. For example, a fifth-grade reading T-score of 45 represents a reading

achievement level that is one-half of a standard deviation lower than the mean for the fifth-grade population represented by the assessed sample of ECLS-K participants. If the same child had a reading T-score of 50 in eighth grade, this would indicate that the child has made up his or her deficit and is reading at a level comparable to the national average.

Appendix A includes tables of subgroup means for the IRT theta (ability) estimates as well as for the IRT scale scores and T-scores. However, because the theta scores may be difficult to use and interpret, except in combination with item parameters, they are not included in the data files.

4.1.3 Item Cluster Scores

Several item cluster scores were reported for the assessments in third and fifth grades. Cluster scores are not reported in the eighth-grade round for the reasons described below; however, descriptions of the scores from previous rounds are included for reference.

The item cluster scores are simple counts of the number right on small subsets of items linked to particular skills. These clusters of items are also included in the broad-range scores described above. Because they are based on very few assessment items, their reliabilities are relatively low.

Reading. The K-1 reading assessment contained three questions assessing children's familiarity with conventions of print. The score for these questions was obtained by counting the number of correct answers for the three items. The print familiarity cluster score is documented in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Psychometric Report for Kindergarten Through the First Grade* (NCES 2002–05) (Rock and Pollack 2002) and is included in the K-1 data files (*Early Childhood Longitudinal Study, Kindergarten–First Grade Public-Use Data Files and Electronic Codebook*, NCES 2002–149) (Tourangeau et al. 2002). These items were not included in the third-grade reading forms because nearly all children had mastered them by the end of first grade.

In addition, a cluster score based on a set of four relatively difficult decoding items was reported for the third- and fifth-grade assessments and is documented in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Psychometric Report for the Fifth Grade* (NCES 2006–036rev) (Pollack, Atkins-Burnett et al. 2005). It is included in the fifth-grade data files

(*Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), User's Manual for the ECLS-K Fifth-Grade Public-Use Data Files and Electronic Codebooks* (NCES 2006–032) (Tourangeau et al. 2006). These decoding words are unlikely to be in most children's everyday vocabulary but could be sounded out phonetically. As mentioned above, no cluster scores are reported for reading in eighth grade.

Science. The 21 routing form items of the fifth-grade science assessment measured a range of basic concepts, with seven questions each in life science, physical science, and earth science. The seven-item clusters administered in the fifth-grade routing test each included the five items tested in the corresponding cluster in third grade that comprised the 15-item third-grade routing form. Due to the small number of test items on the eighth-grade routing form, it was not possible to include all of the cluster items from the three science concepts, and therefore the scores are not reported for the eighth-grade data.

For fifth grade, scores consisting of simple counts of number right for the seven items, as well as for the five-item subsets reported in third grade, were computed for each of the three clusters. Children who omitted more than two items in a cluster were not scored. The items were not selected to have comparable levels of difficulty within each set. For example, the fifth-grade mean of 4.8 for the seven-item life science cluster compared with 4.2 for physical science does not mean in any sense that children were doing better or learning more relative to the domain curriculum in life science compared with physical science. With only five or seven items each, these clusters are not reliable measures of the domain for each content strand. They simply sample a small set of questions of varying difficulty and content within each domain, which may be used for subgroup comparisons.

Mathematics. Item cluster scores were not defined for any round on the mathematics assessments. Although mathematics items were clustered for the proficiency levels (see section 4.1.4), the remaining items were less homogeneous with respect to similarity of content and skill demand, and thus cluster scores were not developed.

4.1.4 **Proficiency Levels**

Proficiency levels provide a means of distinguishing status or gain in specific skills within a content area from the overall achievement measured by the IRT scale scores and T-scores. Clusters of four assessment questions having similar content and difficulty were included at several points along the score scale of the reading and mathematics assessments. Clusters of four items provide a more reliable

assessment of proficiency than do single items because of the possibility of guessing; it is very unlikely that a child who has not mastered a particular skill would be able to guess enough answers correctly to pass a four-item cluster.

The proficiency levels were assumed to follow a Guttman model, that is, a child passing a particular skill level was expected to have mastered all lower levels; a failure should be consistent with nonmastery at higher levels. Only a very small percentage of children in kindergarten through eighth grade had response patterns that did not follow the Guttman model, that is, a failing score at a lower level followed by a pass on a more difficult item cluster. Overall, including all seven rounds of data collection, less than 6 percent of reading response patterns and about 3 percent of mathematics assessment results failed to follow the expected hierarchical pattern. This does not necessarily indicate a different order of learning for these children; since most of the proficiency level items were multiple-choice, many of these reversals may be due to children guessing.

The nine reading and nine mathematics proficiency levels identified in the kindergarten through fifth-grade assessments, and a tenth reading level defined in eighth grade, are described in section 4.3.2. No proficiency scores were computed for the science assessment because the questions did not follow a hierarchical pattern. Two types of scores are reported with respect to the proficiency levels: a single indicator of highest level mastered, and a set of IRT-based probability scores, one for each proficiency level. More information on each of these types of scores is provided below.

4.1.4.1 Highest Proficiency Level Mastered

Mastery of a proficiency level was defined as answering correctly at least three of the four questions in a cluster. This definition results in a very low probability of guessing enough right answers to pass a cluster by chance. The probability varies depending on the guessing parameters (IRT "c" parameters) of the items in each cluster but is generally less than 2 percent. At least two incorrect responses indicated lack of mastery, while omitted items were not counted as either right or wrong. Since the ECLS-K direct cognitive assessment was a two-stage design (where not all children were administered all items), and since more advanced assessment instruments were administered in third, fifth, and eighth grades, children's data did not include all of the assessment items necessary to determine pass/fail for every proficiency level at each round of data collection. The missing information was not missing at random; it depended in part on children being routed to second-stage forms of varying

difficulty within each round, and in part on the range of difficulty of the assessments at the different grade levels. In order to avoid bias due to the non-randomness of the missing proficiency level scores, imputation procedures were undertaken to fill in the missing information.

Pass or fail for each proficiency level was based on actual counts of correct or incorrect responses, if they were present. If too few items were administered or answered to determine mastery of a level, a pass/fail score was imputed based on the remaining proficiency level scores only if they indicated a pattern that was unambiguous. That is, a "fail" might be inferred for a missing level if there were easier cluster(s) that had been failed *and* no higher cluster passed; or a "pass" might be assumed if harder cluster(s) were passed *and* no easier one failed. In the case of ambiguous patterns (e.g., pass, missing, fail for three consecutive levels, where the missing level could legitimately be either a pass or a fail), an additional imputation step was undertaken that relied on information from the child's performance on all of the domain items answered in that round of data collection. IRT-based estimates of the probability of a correct answer were computed for each missing assessment item and used to assign an imputed right or wrong score to the item. These imputed responses were then aggregated in the same manner as actual responses to determine mastery at each of the missing levels.

About 67 percent of the "highest level" scores in reading and 80 percent in mathematics was determined on the basis of item response data alone for the kindergarten through eighth-grade rounds. In eighth grade, the scores determined on the basis of item response data dropped to 19 percent for reading and 47 percent for mathematics, a result of the necessary placement of the proficiency level items on either the low or high second-stage form, based on their estimated difficulty levels. The rest used IRTbased probabilities for some or all of the missing items, since the "missingness" was a consequence of the presence or absence of the necessary items in the test forms children received. The eighth-grade reading assessment did not include any of the proficiency level 7 items; level 8 items appeared only on the low second-stage form; level 9 items were split between the routing and low forms; and level 10 items were in the high form only. Scores were not imputed for missing levels for patterns that included a reversal (e.g., fail, blank, pass) because no resolution of the missing data could result in a consistent hierarchical pattern. In reading, imputation was not possible for 5 percent of the children in fall-kindergarten, 7 percent in spring-kindergarten, 6 percent in fall-first grade, 7 percent in spring-first grade, 7 percent in spring-third grade, 6 percent in spring-fifth grade, and 1 percent in spring-eighth grade. The percents are less in mathematics, with 3 percent in fall-kindergarten, 4 percent in spring-kindergarten, 3 percent in fall-first grade, 3 percent in spring-first grade, 2 percent in spring-third grade, 3 percent in spring-fifth grade, and 3 percent in spring-eighth grade.

Scores in the data file represent the highest level of proficiency mastered by each child at each round of data collection, whether this determination was made by actual item responses alone, or by a combination of item responses and imputed scores. The highest proficiency level mastered implies that children demonstrated mastery of all lower levels and nonmastery of all higher levels. A zero score indicates nonmastery of the lowest proficiency level. Scores were excluded only if the actual or imputed mastery level data resulted in a reversal pattern as defined above. The highest proficiency level mastered scores do not necessarily correspond to an interval scale, so in analyzing the data, they should be treated as ordinal.

The highest proficiency level mastered variable is suitable for both cross-sectional and longitudinal analyses. The user must be aware that, although the proficiency levels have been shown to be hierarchical, there is no claim that they are equal-interval. That is, it would be incorrect to assume that a time 1 to time 2 gain from, for example, level 5 to level 7 is in any sense equivalent to a gain from level 7 to level 9 over the same period. One shouldn't think of the proficiency levels as if they were points gained, but rather as milestones achieved. The milestones are not necessarily equally spaced in terms of the time it normally takes to achieve them, or the "amount of skill" that's required, or any other metric. The most appropriate use of these scores would be in looking at distributions, or percentages achieving some selected level, and examining how these distributions differ for different groups or change over time.

4.1.4.2 Proficiency Probability Scores

Proficiency probability scores are reported for each of the proficiency levels described above, at each round of data collection. The scores estimate the probability of mastery of each level and can take on any value from zero to one. An IRT model was employed to calculate the proficiency probability scores, which indicate the probability that a child would have passed a proficiency level, based on the child's whole set of item responses in the content domain. The item clusters were treated as single items for the purpose of IRT calibration, in order to estimate children's probabilities of mastery of each set of skills. The hierarchical nature of the skill sets justified the use of the IRT model in this way.

The proficiency probability scores differ from the highest level scores in that they can be used to measure gains over time, and from the IRT scale scores in that they target specific sets of skills. The proficiency probability scores can be averaged to produce estimates of mastery rates within population subgroups. These continuous measures can provide a close look at individuals' status and change over time. Gains in probability of mastery at each proficiency level allow researchers to study not only the amount of gain in total scale score points but also where along the score scale different children are making their largest gains in achievement during a particular time interval. For example, subtracting the mathematics level 8 probability at fifth grade from the mathematics level 8 probability at eighth grade would indicate to what extent a child has advanced in mastery of fractions during this time interval. Thus, children's school experiences at selected times can be related to improvements in specific skills.

4.2 Reading Assessment

The eighth-grade reading test emphasized reading comprehension, with all questions based on several reading passages. The reading assessment began with a routing test of 10 items, based on three reading passages, administered in ascending order of difficulty. The score on the routing test was used to select one of two second-stage forms, of varying difficulty, consisting of three (low form) or four (high form) reading passages, each with three to nine associated questions.¹

4.2.1 Samples and Operating Characteristics

Table 4-1 presents sample counts and operating characteristics of the adaptive test forms in reading. Note that the same set of assessment forms was used for rounds 1 through 4, fall-kindergarten through spring-first grade. A new set of assessment forms suitable for third-graders was used in round 5, an additional set for fifth-graders in round 6, and a set for eighth-graders in round 7. The small sample size reported at round 3 in table 4-1 reflects the fact that only a subsample of the fall-first grade longitudinal cohort was assessed at this point in time. Scores were calculated only for children who attempted at least 10 items in the routing test and second-stage form combined. The line labeled "Too few items" refers to the number of children who did not attempt a sufficient number of reading items to generate a reliable score. This number is excluded from the "Total" line, which is the number of scorable tests. Children who lacked sufficient English proficiency to pass the English language screening test, administered in rounds 1 through 4 only, were excluded from the reading assessment in those rounds.² All children were included in the data collections of subsequent rounds.

¹ A change from prior assessment designs was made to reduce the number of second-stage forms in each domain from three to two, with routing to only a low or high second-stage form, eliminating the middle form. For further details, see section 2.1.

² The number of children not passing the screener in rounds 1 through 4 was 1,452, 956, 208, and 338, respectively.

2001-02, 2003-04, and 2000							
Characteristics	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Total	17,630	18,944	5,054	16,340	14,286	11,267	9,225
Too few items	44	19	0	2	134	31	7
Percent taking low form	76	34	21	4	25	26	37
Percent taking middle form	21	47	46	15	56	49	•;•
Percent taking high form	4	19	33	82	19	25	63
Percent perfect score routing test	ς.	1.7	4.9	23.6	3.4	0.1	11.2
Percent perfect score low form	0.0	0.1	0.4	1.6	0.0	0.4	0.1
Percent perfect score middle form	0.0	0.0	0.0	0.0	0.0	0.0	- -
Percent perfect score high form	0.0	0.2	0.0	0.0	0.0	0.5	0.2
Percent less than chance routing test	22.6	3.7	2.1	0.3	0.4	1.3	8.1
Percent less than chance low form	0.9	0.5	0.2	0.6	3.6	0.7	8.7
Percent less than chance middle form	0.5	0.3	0.1	0.1	0.2	0.2	•;
Percent less than chance high form	0.5	1.7	2.3	0.4	0.0	0.0	1.0
† Not applicable.							

Reading assessment: Samples and operating characteristics: Rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02 2003–04 and 2006–07 Table 4-1.

89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. "Too few items" refers to the number of children who did not attempt a sufficient number of reading items (10 items) to generate a reliable score (e.g., 7 children in round 7). Percentages are unweighted. Differences in Ns NOTE: Rounds 1 through 4 used the same set of assessment forms; round 5, round 6, and round 7 forms were different sets developed for third, fifth, and eighth grades, respectively. Approximately

across subjects in the same round are due to children with too few or no responses in the particular subject assessment. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007. The percentages taking the various second-stage forms in reading followed the expected distributions, based on the cut points determined by simulations using field test item parameters and estimates of ability distributions. That is, in round 1 about three-quarters of the children were assigned the low second-stage form based on their routing test performance. In rounds 2 and 3, the largest percentages were assigned the middle-level form. By spring-first grade, round 4, more than three-quarters of the children took the highest level of the second-stage forms. The third- and fifth-grade assessments developed for rounds 5 and 6 were designed to route approximately 50 percent of children to the middle form, with the remaining children about evenly divided between the low and high forms. The cut scores for the eighth-grade assessments (0–5 routed to low form, 6–10 routed to high form) were not designed with target percentages of children, but instead, to expand the range of ability levels expected in eighth grade, using only two second-stage forms.

More important than the routing percentages matching the intended targets is whether the cutting scores succeeded in routing children to a second-stage test of an appropriate level of difficulty. The percentages of perfect and less-than-chance scores are shown in table 4-1. The percentages of perfect scores were all close to zero, with exception of the round 4 and 7 routing tests. Although about 23 percent of children had perfect scores on the routing test in round 4, and 11 percent of children in round 7, the main function of the routing test was to make a proper assignment to the correct second-stage form. The children were then scored on the *combination* of their routing and second-stage items. Since there were virtually no perfect scores on the high-level second stage forms in any round, the perfect routing test scores did not have the potential to create a ceiling effect. Table 4-1 also shows little or no evidence of a floor effect when both first and second stages are combined to compute ability levels and scale scores. While 22.6 percent scored below chance on the routing test in round 1, these children were routed to the low-level second-stage form where more than 99 percent of them were able to respond at or above the chance level. Similarly, although 8.1 percent of children scored below chance on the routing test in round 7, more than 97 percent were able to respond at or above the chance level on the low form, resulting in a small floor effect in eighth grade. A small floor effect also occurred for the least skilled readers in third grade: about 2.5 percent of children were at the chance level or below, with fewer than four correct answers on the routing and second-stage forms combined.

4.2.2 Scores Unique to the Reading Assessment: Proficiency Levels

The following 10 reading proficiency levels were defined for the longitudinal assessments.

Level 1: Letter recognition: identifying upper- and lower-case letters by name;

Level 2: Beginning sounds: associating letters with sounds at the beginning of words;

Level 3: Ending sounds: associating letters with sounds at the end of words;

Level 4: Sight words: recognizing common words by sight;

Level 5: Comprehension of words in context: reading words in context;

Level 6: Literal inference: making inferences using cues that are directly stated with key words in text (for example, recognizing the comparison being made in a simile);

Level 7: Extrapolation: identifying clues used to make inferences, and using background knowledge combined with cues in a sentence to understand use of homonyms;

Level 8: Evaluation: demonstrating understanding of author's craft (how does the author let you know...), and making connections between a problem in the narrative and similar life problems;

Level 9: Evaluating nonfiction: critically evaluating, comparing and contrasting, and understanding the effect of features of expository and biographical texts; and

Level 10: Evaluating complex syntax: evaluating complex syntax and understanding highlevel nuanced vocabulary in biographical text.

The test items on which the proficiency levels were defined were not used in all rounds of data collection, but only in grades for which their difficulty was appropriate. Level 1–3 items appeared only in the K-1 assessments; level 4 in K-1 and third grades; level 5 in all rounds; levels 6–7 in third and fifth grades; level 8 in third, fifth, and eighth grades; level 9 in fifth and eighth grades; and level 10 in eighth grade only. IRT procedures described in sections 3.2 and 5.2 were used to obtain probability estimates for all levels at all rounds so that longitudinal gains in specific skills could be measured.

4.2.3 Reliabilities

Table 4-2 presents reliability statistics for the scores of the eighth-grade reading assessment. K-1, third-, and fifth-grade reliabilities are included in the table for comparison purposes. These reliabilities are unweighted and thus represent the reliabilities of the *actual* assessments, based on the sample administrations, and are not weighted to the population. In general, the more items a test has, and the greater the variance in ability of test takers, the higher the reliability is likely to be.

Reliability measure	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Alpha:							
Routing	.86	.88	.88	.86	.75	.88	.73
Low form	.69	.69	.71	.72	.83	.82	.81
Middle form	.70	.72	.74	.78	.84	.72	†
High form	.90	.88	.93	.92	.79	.76	.75
Split-half:							
Decoding score	ţ	ţ	ţ	Ť	.67	ţ	Ť
Proficiency level 1	.83	.79	.77	.78	÷	+	÷
Proficiency level 2	.76	.76	.73	.70	+	+	÷
Proficiency level 3	.72	.76	.76	.68	+	+	+
Proficiency level 4	.78	.77	.80	.78	.56	+	+
Proficiency level 5	.60	.69	.73	.73	.66	.64	ţ
Proficiency level 6	ť	Ť	ţ	†	.51	.51	ţ
Proficiency level 7	Ť	Ť	t	†	.48	.48	†
Proficiency level 8	Ť	Ť	†	†	.63	.64	.63
Proficiency level 9	Ť	†	†	†	†	.40	.22
Proficiency level 10	†	†	†	ţ	Ť	†	.31
Reliability of theta	.92	.95	.96	.96	.94	.93	.87
Percent agreement of highest proficiency level mastered:							
Percent exact agreement	68	57	57	59	53	52	44
Percent exact+off by 1	98	95	95	96	96	96	89

Table 4-2.Reading assessment reliabilities, rounds 1 through 7: School years 1998–99, 1999–2000,
2001–02, 2003–04, and 2006–07

† Not applicable.

NOTE: Statistics are unweighted. Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. Statistics are unweighted. Statistics for IRT-based scores (percent agreement and reliability of theta) may be different from those in earlier reports due to recalibration of longitudinal scales. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

Internal consistency (alpha) coefficients for eighth grade are comparable to those obtained for K-1, third, and fifth grade. The pattern of alpha coefficients for the routing tests is related to the number of items. For tests with similar characteristics, a larger number of items will result in a higher alpha coefficient. The K-1 reading routing test had 20 items, with 15 items in third grade, 26 in fifth grade, and 10 in eighth grade, and the resulting reliabilities reflect these variations in test length. The alpha coefficients for the second-stage forms in each round tend to be lower than those for the routing test due to the restriction in range of ability among the children sent to the various second-stage forms. Since

the children taking each of these forms are a more homogeneous group with respect to reading performance, the score variances, and thus the alpha coefficients, are lower than they would have been if the whole sample of children had taken each set of items. Only for the high-level K-1 second-stage form, which had much greater variance than did the other forms, did the alpha coefficients approach or exceed .90. The tendency for the restricted variance of the second-stage forms to depress the alpha coefficient was offset in third and eighth grades by the greater number of items in the second-stage forms, relative to the number of items in the routing test. The reliabilities of the second-stage forms are presented for the sake of completeness, although scores on the second-stage forms are not reported separately.

Split-half reliabilities were computed for the scores that are defined by clusters of items: the decoding score and the individual proficiency level scores. Each of these reliabilities is a transformation of the correlation of a subscore based on half of the items in the cluster, with the score based on the other half. The decoding score was present only for third and fifth grades, not for the earlier rounds. In the fifthgrade round, only three of the four items in this cluster were present in the assessment and the fourth item was imputed to produce a score, so a calculation of split-half reliability based on all items was not possible. Split-half reliabilities are presented for the individual proficiency level scores for informational purposes only, because "pass/fail" on the proficiency levels is reported only in the aggregate and not for each level separately. The split-half reliabilities tend to be highest for levels 1–5, where the items are essentially replicates of the same task (e.g., level 1, recognizing letters of the alphabet). Levels 6–10 are based on comprehension of reading passages, where the questions within a level are more loosely related to each other than for the lower levels, resulting in lower internal consistency within levels. Another reason for the lower split-half reliabilities is that, like the alpha coefficients for the second-stage forms, the split-half reliabilities for the more difficult proficiency levels are generally depressed by the restriction in range for the children receiving the harder second-stage forms. Since the children taking the items in each of these levels are a more homogeneous group (relative to the whole sample) with respect to reading performance, the score variances, and thus the reliability coefficients, are lower than they would have been if the whole sample of children had taken each set of items.

The most appropriate estimate of the reliability of the reading assessment is the reliability of the overall IRT ability estimate, theta. This number is based on the variance of repeated estimates of theta and applies to all of the scores derived from the theta estimate, namely, the IRT scale scores, T-scores, and proficiency probabilities. Error variance was estimated as the within-person variance of repeated estimates of theta, averaged over all data cases. The ratio of this number to the total variance (between-person variance of the posterior mean) is the estimated proportion of total variance that is error variance,

and 1 minus the proportion of error variance is the estimate of true variance that is reported as the reliability of theta. This reliability index differs from the information function primarily in that it is a single estimate for the whole set of scores, rather than a function evaluated at each point along the continuum. This is the most appropriate estimate of the reliability of the assessment since it reflects the internal consistency of performance on the combined first- and second-stage sections, and for the full range of variance found in the sample as a whole. The reliability of theta applies to the scale scores and proficiency probabilities as well, since these scores are nonlinear transformations of the thetas that do not affect rank orderings.

It was not possible to apply standard measures of reliability to the "highest proficiency mastered" score, for the following reasons. The score is not a set of items replicating the same or similar tasks, so an internal consistency measure such as split-half reliability or alpha coefficient cannot be computed. Nor can the reliability be evaluated based on the variance of repeated estimates of overall ability that was appropriate for the IRT-based scores.

The definition of reliability—consistency of measurement under different circumstances suggested an appropriate way to assess the reliability of the "highest proficiency level mastered" score. The score denoting the highest level mastered reduces the series of pass/fail scores on the hierarchical set of proficiency levels to a single score. For example, a child demonstrating mastery of the first five reading levels but not the remaining five would be said to have a "highest proficiency mastered" score of five. The question to be answered by a reliability estimate is how likely it would be that the same highest level score would be obtained under other circumstances. In this case, the other circumstances available are not a parallel set of items, but two different methods of arriving at the score. A child's highest level mastered could be determined on the basis of actual item response data alone for only 19 percent of the eighthgrade sample (see section 4.1.4.1). Alternatively, IRT ability estimates and item parameters could be used to generate pass/fail scores, and the composite highest level scores, for these same children. The percentage of cases for which these two different methodologies result in identical or adjacent "highest level mastered" scores can be considered to be a reliability estimate.

4.2.4 Score Statistics

Table 4-3 presents weighted reading scale score means for each round. These scores are estimates of the number of correct answers that would have been expected if at every round each child

had been given all of the 212 test items. (Four additional items in third and fifth grades, consisting of difficult decoding words, were used for the purpose of calibrating IRT ability for these rounds, but deleted from the score scale to bring the representation of content strands more closely into alignment with the framework specifications.) One tested item from fifth grade and two from eighth grade were deleted from scoring due to differential item functioning (DIF) (see section 4.3.5). The IRT procedures described earlier allowed the ability estimates to be computed based on the subset of questions actually administered to each child at each round. Scale scores could then be computed based on the whole item pool. As the assessments progressed from kindergarten through eighth grade, more and more of the test items relied on comprehension of reading passages. Inspection of the reading scale score means by round shows an accelerated rate of growth between fall and spring of first grade, round 3 to round 4, and much larger gains between first and third grade, round 4 to round 5. These gains correspond to the times when children would be mastering basic technical reading skills, and then later, acquiring the ability to derive meaning from what they read. The greater variability in reading performance in the later rounds, compared with kindergarten and fall first grade, can be interpreted as an increase in the reading skills gap between low and high achievers. Weighted score statistics for all reading scores, with breakdowns by population subgroups, are presented in appendix A.

Table 4-3. Reading assessment scale score means and standard deviations, rounds 1 through 7: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

Item	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Scale score							
Mean	34.8	45.7	52.3	75.9	124.7	147.1	166.5
Standard deviation	9.8	13.5	17.7	23.8	28.4	27.5	29.3

NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3A5W0, C4A3W0, C5CW0, C6CW0, C7CW0). Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. Estimates for kindergarten through eighth grade have been put on a common scale to support comparisons, so statistics differ from those in earlier reports. The range of values is 0–212. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

4.2.5 Standard Errors

The method used to calculate the standard errors is described in section 3.2.3. The standard error of theta (θ) for each child is calculated from the sum of item information functions for all items answered by that child. Table 4-4 lists the mean thetas and mean standard errors by round. As expected, the standard errors are lower in the middle range of the ability distribution, in the middle rounds, and

greater at the tails, or at the earlier and later rounds because the distribution of item difficulties is sparse in ranges where few test takers were expected.

Table 4-4.	Reading assessment mean theta score and mean standard error, rounds 1 through 7: School
	years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

Item	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Mean theta	-1.30	-0.72	-0.49	0.12	0.79	1.05	1.34
Mean standard error of	0.20	0.17	0.15	0.12	0.12	0.11	0.20
theta							

NOTE: Table estimates are unweighted. Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. Estimates for kindergarten through eighth grade have been put on a common scale to support comparisons, so statistics differ from those in earlier reports.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

4.2.6 Differential Item Functioning

Section 3.3 explains the DIF procedures used for identifying test items that perform differentially for population subgroups. An adequate number of child responses (minimum of 300 per item) were available to perform DIF analyses on *all* reading items for the male/female and White/Black contrasts. Due to an insufficient number of responses on a portion of the reading items, 78 percent and 67 percent of the items were analyzed for the White/Hispanic and White/Asian contrasts, respectively. Table 4-5 summarizes the results of the DIF analysis of the eighth-grade reading items. Two items exhibited DIF³ against females, one against Black children, while another favored Black children. The two items exhibiting DIF against females were dropped from calibration and scoring. Although these items had been previously administered in NAEP and did not exhibit DIF in those administrations, the fairness committee review recommended dropping these items. The other item, favoring Black children, was retained, since the item content did not appear to favor one group or another, based on the fairness review.

Table 4-5. Reading assessment: Differential item functioning, eighth grade: School year 2006–07

Reference group:	Male	White	White	White	White
Focal group:	Female	Black	Hispanic	Asian	Other
Number of DIF items favoring focal group	0	1	0	0	0
Number of DIF items favoring reference group	2	1	0	0	0

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

³ DIF refers to the combined finding of Mantel-Haenzsel C-level DIF and P-DIF greater than 10 percent (see section 3.3).

It should be kept in mind that there were 45 reading items in the eighth-grade reading assessment forms and five sets of comparison groups resulting in several hundred comparisons. The large number of contrasts evaluated means that chance alone could result in statistically significant differences for a few items even where no differential functioning actually exists.

4.3 Mathematics Assessment

The eighth-grade mathematics framework specifications differed from those for third and fifth grades, with an increased emphasis on the more difficult content strands of geometry and spatial sense; data analysis, statistics, and probability; and patterns, algebra, and functions. The transition from the easier content strands of measurement and number sense, properties, and operations, to the more difficult content strands is directly related to the curriculum standards differences between eighth grade and the elementary grades. Children began the mathematics assessment with a routing test of 10 items. The score on the routing test was used to select one of two second-stage forms of varying difficulty, each consisting of 20 items.

4.3.1 Samples and Operating Characteristics

Table 4-6 presents sample counts and operating characteristics of the adaptive test forms in mathematics. Note that the same set of assessment forms was used for rounds 1 through 4, fall-kindergarten through spring-first grade. A Spanish translation of the mathematics assessment was administered in kindergarten and first grade to children who were Spanish speakers and whose English language fluency was not sufficiently advanced to take the assessments in English. Children who lacked English language fluency but were not Spanish speakers were excluded from the mathematics assessment in those rounds (415 children in round 1, 229 children in round 2, 34 children in round 3, and 37 children in round 4). More advanced sets of assessment forms, entirely in English, were developed for third, fifth, and eighth grades. Scores were calculated only for children who attempted at least 10 items in the routing test and second-stage form combined.

The eighth-grade assessment developed for round 7 was designed to route children to a second-stage form with an appropriate difficulty level and did not target percentages of children per form. However, the data collection resulted in an approximate split of the sample between the low and high forms for mathematics. Again, the important point here is not matching routing percentages, but selecting the test form that best matches each child's ability level. The cutting points for the routing test were selected to minimize floor and ceiling effects rather than to match target distributions. The percentages of perfect and less-than-chance scores are shown in table 4-6. The percentages of perfect scores were all close to zero with exception of the rounds 4 and 7 routing tests. Although about 8 percent of children had perfect scores on the routing test in round 4, and 12 percent of children in round 7, the main function of the routing test was to make a proper assignment to the correct second-stage form. The children were then scored on the *combination* of their routing and second-stage items. Since there was no ceiling effect problem (virtually no perfect scores) in the high-level, second-stage form for rounds 1 through 6, and only about 1.4 percent in round 7, the perfect routing test scores did not create a ceiling effect for the test as a whole. Table 4-6 also shows little or no evidence of a floor effect when both first and second stages are combined to compute ability levels and scale scores. While 22.6 percent scored below chance on the routing test in round 1, these children were routed to the low-level, second-stage form where more than 99 percent of them were able to respond at or above the chance level. Similarly, although 9.7 percent of children scored below chance on the routing test in round 7, over 97 percent were able to respond at or above the chance level on the low form, resulting in a near negligible floor effect in grade 8.

99–2000,	
998-99, 19	
years 1	
7: School	
l through 7	
rounds 1	
characteristics,	
perating	
s and c	
Samples	20-900
hematics assessment:	1–02, 2003–04, and 20
Mat	200
Table 4-6.	

Characteristics	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Total	18,641	19,657	5,226	16,647	14,380	11,276	9,285
Too few items	21	15	0	2	29	22	5
Percent taking low form	77	43	26	L	29	36	52
Percent taking middle form	17	31	29	14	37	34	•;•-
Percent taking high form	9	26	45	62	33	30	48
Percent perfect score routing test	0.1	0.4	1.5	7.9	1.6	1.8	11.8
Percent perfect score low form	0.1	0.4	1.0	2.5	0.0	0.6	0.4
Percent perfect score middle form	0.0	0.0	0.0	0.3	0.1	0.0	•;
Percent perfect score high form	0.0	0.0	0.0	0.1	0.0	0.2	1.4
Percent less than chance routing test	15.3	3.1	1.6	0.3	1.3	1.5	9.7
Percent less than chance low form	0.0	0.3	0.1	0.3	0.3	0.4	2.6
Percent less than chance middle form	0.1	0.0	0.0	0.0	0.1	0.1	•;
Percent less than chance high form	0.1	0.0	0.0	0.0	0.1	0.7	0.2
† Not applicable.							

NOTE: Rounds 1 through 4 used the same set of assessment forms; round 5, round 6, and round 7 forms were different sets developed for third, fifth, and eighth grades, respectively. Some children in rounds 1 through 4 received a Spanish translation of the mathematics assessment; in rounds 5, 6, and 7, all assessments were in English. Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. "Too few items" refers to the number of children who did not attempt a sufficient number of mathematics items to generate a reliable score. Percentages are unweighted. Differences in Ns across subjects in the same round are due to children with too few or no responses in the particular subject assessment.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

4.3.2 Scores Unique to the Mathematics Assessment: Proficiency Levels

The following nine mathematics proficiency levels were defined for the longitudinal assessments.

Level 1: Number and shape: identifying some one-digit numerals, recognizing geometric shapes, and one-to-one counting of up to 10 objects.

Level 2: Relative size: reading all single-digit numerals, counting beyond 10, recognizing a sequence of patterns, and using nonstandard units of length to compare objects.

Level 3: Ordinality, sequence: reading two-digit numerals, recognizing the next number in a sequence, identifying the ordinal position of an object, and solving a simple word problem.

Level 4: Addition/subtraction: solving simple addition and subtraction problems.

Level 5: Multiplication/division: solving simple multiplication and division problems and recognizing more complex number patterns.

Level 6: Place value: demonstrating understanding of place value in integers to the hundreds place.

Level 7: Rate and measurement: using knowledge of measurement and rate to solve word problems.

Level 8: Fractions: demonstrating understanding of the concept of fractional parts.

Level 9: Area and volume: solving word problems involving area and volume, including change of units of measurement.

As was the case for reading, the test items on which the mathematics proficiency levels were defined were not used in all rounds of data collection, but only in grades for which their difficulty was appropriate. Level 1–3 items appeared only in the K-1 assessments, level 4 in K-1 and third grades, level 5 in K-1 through fifth grade, level 6 in third and fifth grades, and levels 7–9 in third through eighth grades. IRT procedures described in sections 3.2 and 5.2 were used to obtain probability estimates for all levels at all rounds so that longitudinal gains in specific skills could be measured.

4.3.3 Reliabilities

Table 4-7 presents unweighted reliability statistics for the scores of the eighth-grade mathematics assessment. K-1, third-, and fifth-grade (unweighted) reliabilities are included in the table for comparison purposes.

Table 4-7.	Mathematics assessment reliabilities, rounds 1 through 7: School years 1998–99, 1999–2000,
	2001–02, 2003–04, and 2006–07

Reliability measure	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Alpha:							
Routing	.78	.81	.83	.80	.86	.88	.76
Low form	.70	.66	.66	.71	.77	.78	.82
Middle form	.66	.67	.66	.66	.72	.58	ţ
High form	.80	.80	.83	.82	.73	.75	.76
Split-half:							
Proficiency level 1	.41	.27	.26	.26	†	Ť	†
Proficiency level 2	.58	.49	.51	.32	†	Ť	†
Proficiency level 3	.63	.66	.67	.59	†	†	†
Proficiency level 4	.54	.63	.66	.63	.43	†	t
Proficiency level 5	.46	.53	.61	.65	.67	.64	†
Proficiency level 6	†	Ť	Ť	†	†	.78	†
Proficiency level 7	†	†	Ť	†	.42	.68	.63
Proficiency level 8	†	Ť	Ť	†	†	.56	.72
Proficiency level 9	Ť	Ť	Ť	Ť	Ť	.48	.54
Reliability of theta	.91	.93	.94	.94	.95	.95	.92
Percent agreement of highest proficiency level mastered:							
Percent exact agreement	59	56	56	57	55	57	61
Percent exact+off by 1	97	97	97	98	96	96	98

† Not applicable.

NOTE: Statistics are unweighted. Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. The four test items for mathematics proficiency level 6 did not all appear in the same test form in third grade, so no complete data cases were available for evaluation of split-half reliability. Statistics for IRT-based scores (percent agreement and reliability of theta) may be different from those in earlier reports due to recalibration of longitudinal scales. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004, and spring 2007.

All other things being equal (e.g., the psychometric quality of test items), internal consistency coefficients tend to be higher when tests are longer and lower when the ability range of the

test takers is restricted. The internal consistency (alpha) coefficients for the third- and fifth-grade mathematics routing tests were slightly higher than that of the K-1 forms, probably partly due to a slightly longer test (17 and 18 items in third and fifth grades, respectively, vs. 16 items in K-1), and partly because of greater variability in the mathematics achievement of third- and fifth-graders compared with earlier rounds. Similarly, the alpha reliability is lower for the eighth grade routing form, with only 10 items administered. The eighth-grade second-stage mathematics forms have alpha coefficients similar to those of the routing test because the greater number of items in each form offsets the reductions in variance that would be expected as a consequence of the restriction of range of ability within each form. While the K-1 high second-stage form had many more items than the other forms (31 items, compared with 18 and 23 for the low and middle K-1 forms, respectively) and thus a higher reliability coefficient, the third- and fifth-grade tests all had about the same number of items in each second stage form, and similar alphas. The reliabilities of the second-stage forms are presented for the sake of completeness, although scores on the second-stage forms are not reported separately.

Split-half reliabilities are shown in the table for the items present at each round. There is no split-half reliability presented for proficiency level 6 in third grade because the items on which it is based did not all appear in the same test form, so no complete data cases were available for evaluation of the reliability. The kindergarten and first-grade split-half reliabilities for levels 1 through 5 were substantially lower than for the corresponding levels in the reading test. While the sets of reading items in each of the lowest proficiency levels were essentially replicates of the same task, the mathematics sets in the early years were not as homogeneous with respect to content and skill demands. The greater heterogeneity for the mathematics sets may have contributed to their lower split-half reliabilities. Both alpha coefficients and split-half reliabilities tend to be underestimates of "true" reliability, and this tendency may be accentuated by greater diversity of content. The relatively low split-half reliabilities for mathematics proficiency levels 8 and 9 in fifth grade and level 9 in eighth grade are a consequence of their placement only in the high-level form, resulting in restriction in the range of ability of children taking these items.

The reliabilities of the mathematics theta scores were in the .90s for all rounds. The reliability of theta applies to the scale scores and proficiency probabilities as well, since these scores are nonlinear transformations of the thetas that do not affect rank orderings.

The percentages of agreement between methods in determining the highest mathematics proficiency level mastered were comparable to those for reading, both for percentage of exact agreement and percentage of agreement within one level. The greater homogeneity of the reading items for the low compared with high proficiency levels resulted in percent agreement of highest level that tended to go down in the later rounds. Conversely, percent agreement for mathematics, with greater homogeneity in the *later* rounds, tended to go up. See section 4.2.3 for a detailed explanation of how this score was computed and evaluated.

4.3.4 Score Statistics

The weighted scale score means presented in table 4-8 represent estimates of the number of correct answers that would have been expected if each child was administered and had responded to all of the 174 mathematics items in the pool; that is, all items that appeared in the K-1, third-grade, fifth-grade, and/or eighth-grade test forms. The greatest gains are observed between rounds 4 and 5, spring-first grade to spring-third grade. The variance in mathematics achievement increased markedly for each successive round from fall-kindergarten through third grade, leveling off in fifth and eighth grades. Weighted score statistics for the mathematics scores and breakdowns by population subgroups are presented in appendix A.

Table 4-8.Mathematics assessment scale score means and standard deviations, rounds 1 through 7:
School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

Item	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Scale score							
Mean	25.5	35.6	42.7	60.3	97.0	120.6	138.7
Standard deviation	8.9	11.8	14.3	18.2	25.0	25.9	23.6

NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3A5W0, C4A3W0, C5CW0, C6CW0, C7CW0). Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. Estimates for kindergarten through eighth grade have been put on a common scale to support comparisons, so statistics differ from those in earlier reports. The range of values is 0–174. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

Three geometry items that had weak statistics in the third-grade assessment were satisfactory in the fifth-grade round. Data for these items from both rounds were pooled, and the items were included in the longitudinal scale.

4.3.5 Standard Errors

As described in section 3.2.3, the standard errors of theta are calculated from the sum of item information functions for each item *answered by* each child. Table 4-9 lists the mean theta values and mean standard errors by round. The standard errors are roughly the same magnitude as those found in the reading assessment.

Table 4-9.Mathematics assessment mean theta score and mean standard error, rounds 1 through 7:
School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

Item	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Mean theta	-1.17	-0.69	-0.44	0.06	0.72	1.12	1.48
Mean standard error of	0.20	0.18	0.17	0.15	0.13	0.13	0.18
theta							

NOTE: Table estimates are unweighted. Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. Estimates for kindergarten through eighth grade have been put on a common scale to support comparisons, so statistics differ from those in earlier reports.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

4.3.6 Differential Item Functioning

Table 4-10 presents counts of the DIF⁴ items for the eighth-grade mathematics forms. An acceptable number of child responses (300) per item were available to perform DIF analyses on *all* mathematics items for the male/female and White/Black contrasts. Due to an insufficient number of responses on a portion of the mathematics items, 75 percent and 73 percent of the items were analyzed for the White/Hispanic and White/Asian contrasts, respectively. Only one mathematics item exhibited DIF favoring Asian children. This mathematics item was reviewed for fairness and found to be relevant to the construct being measured by the assessment and was retained for scoring. See section 3.3 for an explanation of DIF procedures.

⁴ As before, DIF refers to the combined finding of Mantel-Haenzsel C-level DIF and P-DIF greater than 10 percent (see section 3.3).

	Male	White	White	White	White
Reference group: Focal group:	Female	Black	Hispanic	Asian	Other
Number of DIF items favoring focal group	0	0	0	1	0
Number of DIF items favoring reference group	0	0	0	0	0

Table 4-10. Mathematics assessment: Differential item functioning, eighth grade: School year 2006–07

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

4.4 Science Assessment

The eighth-grade science assessment consisted of a 10-item routing test followed by lowand high-difficulty, second-stage forms, each containing 17 items. Content of the science questions was approximately equally divided among life science, earth science, and physical science strands, with a slight emphasis on earth science. The science assessment was first added to the ECLS-K cognitive battery in third grade; thus the longitudinal score scale spans only third to eighth grades.

4.4.1 Samples and Operating Characteristics

Table 4-11 presents sample counts and operating characteristics of the eighth-grade science forms. Scores were calculated only for children who attempted at least 10 items.

Far fewer children were routed to the low second-stage form, and more to the high form, than had been anticipated based on field test results. As noted above for reading and mathematics, the success of the two-stage procedure is demonstrated by the relative absence of serious floor and ceiling effects. Less than 1 percent of children received a perfect score on the routing, plus second-stage items combined. The percentage of "less than chance" scores in the table is problematic for the children taking the fifth-grade low form. Although a substantial number of children received less than chance scores on the middle and high fifth-grade second-stage forms, when their item responses were combined with routing test responses, none were below chance. However, about 5 percent of children routed to the low second-stage form, or about half of 1 percent of the sample, found the science assessment too difficult overall. Similarly, in round 7, about 5.8 percent of children routed to the low second-stage form found the science assessment too difficult overall, which is about 1 percent of the sample.

Characteristics	Round 5	Round 6	Round 7
Total	14,357	11,273	9,304
Too few items	41	25	3
Percent taking low form	29	13	20
Percent taking middle form	50	41	÷
Percent taking high form	21	46	80
Percent perfect score routing test	1.5	1.1	12.5
Percent perfect score low form	0.3	0.0	0.4
Percent perfect score middle form	0.0	0.1	ť
Percent perfect score high form	0.0	0.1	1.4
Percent less than chance routing test	4.7	1.1	5.3
Percent less than chance low form	1.7	4.9	6.8
Percent less than chance middle form	0.6	4.0	t
Percent less than chance high form	0.8	9.9	3.7

Table 4-11.Science assessment: Samples and operating characteristics, rounds 5 through 7: School
years 2001–02, 2003–04, and 2006–07

† Not applicable.

NOTE: No science assessment was conducted in rounds 1 through 4. The round 5, round 6, and round 7 assessment forms were developed for third, fifth, and eighth grades, respectively. Percentages are unweighted. Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. "Too few items" refers to the number of children who did not attempt a sufficient number of science items to generate a reliable score. Differences in Ns across subjects in the same round are due to children with too few or no responses in the particular subject assessment.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002, spring 2004, and spring 2007.

4.4.2 Reliabilities

Table 4-12 presents unweighted reliability coefficients for the third-, fifth-, and eighth-grade science assessments. Reliabilities for the eighth-grade routing and second-stage forms are similar, and are somewhat lower than those for reading and mathematics. The competing effects on the second-stage forms of the increased number of items, with restricted variance, results in reliabilities comparable to those of the routing form. Alpha coefficients for science forms are generally somewhat lower than those for reading and mathematics are generally somewhat lower than those for reading and mathematics because the science assessment had fewer items in the second-stage forms. This is especially true for the fifth-grade science assessment, in which the routing test was lengthened to 21 items (from 15 in third grade) and the second-stage forms shortened to 14 to 17 items (from 20 in third grade) in order that the items designated for the three science cluster scores would be administered to all children. As a result, the alpha coefficient is higher for the routing test and lower for the second-stage

forms than was the case in third grade. Conversely, in eighth grade, a short routing test and longer second-stage test, offset by the restricted range of ability in the second-stage forms, resulted in nearly identical alpha coefficients for the test sections. Scores for the second-stage forms are not reported separately.

The split-half-reliabilities for the fifth-grade science clusters were somewhat lower than for the decoding cluster in the reading test (.67). Similarly, the reliability of the IRT theta based on all assessment items, and the scores derived from it, is lower than the mid .90s found in reading and mathematics due to the greater diversity of content in the science domain.

Table 4-12.Science assessment reliabilities, rounds 5 through 7: School years 2001–02, 2003–04, and
2006–07

Reliability measure	Round 5	Round 6	Round 7
Alpha:			
Routing	.75	.79	.70
Low form	.70	.54	.68
Middle form	.61	.63	Ť
High form	.60	.48	.70
Split-half:			
Life Science 5-item cluster	.59	.59	Ť
Physical Science 5-item cluster	.49	.41	Ť
Earth Science 5-item cluster	.46	.52	Ť
Life Science 7-item cluster	ť	.64	Ť
Physical Science 7-item cluster	÷	.43	Ť
Earth Science 7-item cluster	Ť	.62	ť
Reliability of theta	.88	.87	.84

† Not applicable.

NOTE: Statistics are unweighted. Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. The 5-item and 7-item clusters scored in rounds 5 and 6 did not appear in the round 7 science test forms (see section 4.1.3).

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002, and spring 2004, and spring 2007.

4.4.3 Score Statistics

Third-, fifth-, and eighth-grade weighted science scale score statistics are presented in table 4-13 and represent the number of correct answers that would have been expected if each child had been given all of the 111 items in all of the test forms. Despite the diversity of content in the assessment, all items had acceptable fit to the IRT model. Weighted score statistics for all science scores and breakdowns by population subgroups are presented in appendix A.

Table 4-13.Science scale score mean and standard deviation, rounds 5 through 7: School years 2001–02,
2003–04, and 2006–07

Item	Round 5	Round 6	Round 7
Scale score			
Mean	49.3	62.8	82.2
Standard deviation	15.1	16.2	17.3

NOTE: Table estimates are based on cross-sectional weights within each round (C5CW0, C6CW0, C7CW0). Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. Estimates for third through eighth grade have been put on a common scale to support comparisons, so statistics differ from those in earlier reports. The range of values is 0-111.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002, spring 2004, and spring 2007.

4.4.4 Standard Errors

Table 4-14 lists the mean theta values and mean standard errors by round. The standard errors are higher than those found in the reading and mathematics assessments. The reasons for this are two-fold. First, as described in section 3.2.3, the standard errors of theta are calculated from the sum of item information functions for each item *answered by* each child. The science assessments generally contained fewer items per round, and thus the test information function was a sum of a smaller number of items. Second, the diverse content administered in the science assessment resulted in discrimination parameters that were generally lower than for reading and mathematics, contributing to the higher standard errors of the ability estimates.

Table 4-14.	Science mean theta score and mean standard error, rounds 5 through 7: School years 2001–02,
	2003–04, and 2006–07

Item	Round 5	Round 6	Round 7
Mean theta	-0.62	-0.01	0.98
Mean standard error of theta	0.37	0.38	0.53

NOTE: Table estimates are unweighted. Approximately 89 percent of the round 7 children were in eighth grade during the 2006–07 school year, 9 percent were in seventh grade, and about 2 percent were in sixth or other grades. Estimates for third through eighth grade have been put on a common scale to support comparisons, so statistics differ from those in earlier reports.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002, spring 2004, and spring 2007.

4.4.5 Differential Item Functioning

Table 4-15 summarizes the results of the DIF analysis of the eighth-grade science items. An adequate number of child responses (300) per item were available to perform DIF analyses on *all* science items for the male/female, White/Black, and White/Hispanic contrasts. A portion of the science items had an insufficient number of responses, resulting in 75 percent of the items analyzed for the White/Asian contrast. Only one item was identified as having DIF,⁵ and it favored the focal group (Black children). This item was reviewed and found to be relevant to the construct being measured by the assessment, so it was retained in the scoring procedures.

Table 4-15.	Science assessment:	Differential ite	m functioning.	eighth	grade: School	vear 2006–07
				- 0 -	0	J

Reference group:	Male	White	White	White	White
Focal group:	Female	Black	Hispanic	Asian	Other
Number of DIF items favoring focal group	0	0	0	0	0
Number of DIF items favoring reference group	0	1	0	0	0

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002 and spring 2007.

Section 3.3 explains the DIF procedures used for identifying test items that perform differentially for population subgroups.

⁵ As before, DIF refers to the combined finding of Mantel-Haenzsel C-level DIF and P-DIF greater than 10 percent (see section 3.3).

5. DIRECT COGNITIVE ASSESSMENTS: LONGITUDINAL MEASUREMENT

The study of the relationships between children's school experiences and their gains in academic skills requires accurate measurements of achievement on scales that can be linked across years. This chapter discusses issues in the longitudinal measurement of the reading and mathematics skills of ECLS-K children from fall-kindergarten through spring-eighth grade, and of science skills from spring-third grade to spring-eighth grade. The development of the longitudinal scales, including analysis of common items, is described. Evidence supporting the validity of the measures is presented. The final section of the chapter focuses on applications: choosing the appropriate scores for analysis and interpreting gain statistics.

5.1 Development of the K-1-3-5-8 Longitudinal Scale

The longitudinal scales necessary for measuring gain over time were developed by pooling the four rounds of kindergarten and first-grade data with the data from the ECLS-K third, fifth, and eighth grades. Data from a small sample of second-graders were included to support the development of the scales by bridging the anticipated gap in ability between first and third grades. The link between the assessment forms used in different rounds relied on the presence of common items shared by successive test forms.

The scale scores for kindergarten and first grade were based on the pool of items used in the test forms administered in those grades. Items were added to the pools as each successive round of data was collected: a supplementary set of reading items in first grade, and new assessment forms for the third-, fifth-, and eighth-grade rounds. Thus the kindergarten reading scale scores were estimates based on a pool of 72 items, with the pool expanding to 92 items for kindergarten and first grade combined, and to 154, 186, and then 212 items as the third-, fifth-, and eighth-grade assessments were added. Each time the item pool was expanded, scores were recalibrated for *all* rounds to make longitudinal comparisons possible. Each recalibration of the scale score represents the estimated number right on a larger and larger set of items that includes all of the items in the current round as well as all administered in previous rounds. As a result, the scale score for the *same* child in the *same* grade changes each time a new set of test items is incorporated and the scale on which the score is based is expanded.
5.1.1 Second-Grade Bridge Study

Chapter 2, section 2.1.5 of the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade (NCES 2005–062) (Pollack, Rock et al. 2005) documents the gap in ability levels that was anticipated due to the absence of the second-grade data collection from the longitudinal design. Without any second-grade data, the accuracy of measurement of cognitive gains from first to third grade might have been compromised. Many of the cognitive test items linking the kindergarten through first-grade (K-1) assessments with the third-grade forms were too hard for most first-graders and too easy for most third-graders. Stable estimates of item parameters necessary for establishing the longitudinal scale require that there be substantial numbers of test takers whose ability levels match the difficulty of the linking items. These test takers did not need to be part of the ECLS-K longitudinal cohort. They needed only to have ability levels in the range where the ECLS-K longitudinal sample data might be sparse, and to take sets of cognitive test items that included the items designed to link the first- and third-grade rounds. Section 5.1 of the above-referenced report describes in detail the collection of reading and mathematics data for a sample of approximately 900 second-graders who were not part of the ECLS-K longitudinal sample. It documents the characteristics of the second-grade bridge sample and shows how the data were used to supplement the longitudinal sample data in establishing vertical scales for measurement of gain. Since the purpose of the bridge sample was to obtain data on the performance of the assessment items, rather than track the progress of the children themselves, their assessment scores are not included in released data files.

The absence of a fourth-grade round of data collection in ECLS-K also represented a potential gap in abilities that could affect the longitudinal scale. However, examination of field test results for fourth- and fifth-graders compared with third-graders showed that sufficient overlap of ability levels from third to fifth grade existed, and that a fourth-grade bridge sample was unnecessary. Similarly, examination of the ECLS-K eighth-grade field test results showed that there was adequate overlap between the fifth-grade items and those intended for eighth grade such that a bridge sample for sixth and seventh grades was also unnecessary.

5.1.2 Evaluating Common Items

Linking score scales across grades required not only overlapping ability distributions, but also overlapping test forms. The longitudinal score scales relied on common items that were present in more than one set of assessment forms. These common items permitted the development of a vertical scale suitable for measuring gains in the elementary through middle school years. Table 5-1 shows the number of items in each subject area shared by more than one set of assessment forms, as well as the number that appeared in only one set. Within rounds, the score scale was supported by items taken by all children within the round (the 10 to 25 items on the routing tests), as well as smaller numbers of items overlapping adjacent forms.

Table 5-1.Counts of common items, separate items, and total items in item pools:
School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

Assessment versions	Reading	Mathematics	Science
Total item pool	212	174	111
Common items (total)	81	49	38
K-1 and third grade	13	9	Ť
First-grade reading supplement and fifth grade	2	Ť	Ť
Third and fifth grade	40	17	21
K-1 (incl. first-grade supplement), third and fifth grade	9	4	Ť
Fifth and eighth grade	12	9	11
Third, fifth and eighth grade	5	10	6
Separate items (total)	131	125	73
K-1 only (including first-grade supplement)	68	50	ţ
Third grade only	16	34	35
Fifth grade only	21	20	19
Eighth grade only	26	21	19

† Not applicable.

NOTE: Four additional reading items were used in calibration of abilities but deleted from the scale scores to bring the content representation into closer alignment with framework specifications. Two of these items that were shared appeared in both the third-grade and fifth-grade assessment forms, the other two in fifth grade only. Science was not tested in kindergarten or first grade.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

The first step in developing the longitudinal scale was evaluating the functioning of the common items at different time points. Although the content and presentation of each of the common items were identical in the four versions of the assessments (K-1, third grade, fifth grade, and eighth grade), it was still possible for the items to function differently. Of course, it would be expected that performance on the items would improve as children advance through school and gain skills, and gains in the probability of a correct answer would be observed. However, the *relative* difficulty of items in the context of the whole assessment should be maintained for the common items used to anchor the scale. For

example, an item "X" based on content that had not yet been introduced could, in fifth grade, be the hardest item in the assessment and could be found to be much more difficult than a particular set of computation items "Y." By eighth grade, when children could have had extensive practice in the skills measured by "X," it could become much *easier* than the *same* set of "Y" computations. Such an item, showing a large difference in *relative* difficulty over time, should not be treated as a common item for the purpose of estimating gains.

In order to assess the common *functioning* of the overlapping reading, mathematics, and science items, preliminary estimates of item response theory (IRT) item and ability parameters were obtained, using all items in the K-1, third-grade, fifth-grade, and eighth-grade assessment forms. For this purpose, each common item was initially assumed to be common functioning, and then this assumption was tested as follows. Responses for each of the common items were pooled for all rounds, and a single set of item parameters was estimated for each. Then the *actual* performance on the common items in each round was compared with performance *predicted* by the IRT item and ability parameters, in order to identify discrepancies that would indicate differential functioning for any items.

Tables 5-2 through 5-4 compare the actual with the predicted proportion correct for each of the reading, mathematics, and science items used in more than one assessment version, based on the children who answered each of the items in each round of data collection. Note that the comparisons of observed vs. predicted percent correct for each question can be carried out *only for children who answered the question*. Many questions appeared in only one or two second-stage forms within a grade, or after a discontinue point in the routing test. Thus most of the items were answered by only a subset of children tested in each round. The statistics shown in tables 5-2 through 5-4 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.

For almost all of the items, the difference between the observed and predicted percent correct was very small (either positive or negative), indicating common functioning of the items across time periods and good fit to the IRT model. Only one item common to the K-1 and third-grade mathematics assessments had a sufficiently large discrepancy in actual compared with predicted proportion correct to warrant separate calibration. This item was deleted from the common item list used for anchoring the scale, but retained for each (K-1 and third-grade) assessment form, with separate sets of item parameters. No non-common-functioning items were found in the reading and science assessments.

	Used in							
Item	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
RUNS	K1,3	-0.01	0.00	0.00	0.03	-0.01	ţ	†
WENT	K1,3	0.00	0.00	-0.01	0.02	-0.01	+	†
DOWN	K1,3	0.00	-0.03	-0.01	0.05	-0.01	+	+
JEEP	K1,3	0.01	0.00	-0.01	0.02	-0.06	+	+
QUIET	K1,3	+	ţ	0.01	0.00	-0.01	+	+
RAGE	K1,3	0.05	0.01	-0.01	-0.01	0.03	+	+
TOIL	K1,3	0.06	0.02	0.00	-0.01	0.04	+	+
CORNER	K1,3	0.03	0.00	-0.01	-0.01	0.04	+	+
REQUIRE	K1,3	+	ţ	0.01	-0.02	0.02	+	+
CAPTURE	K1,3	0.04	0.00	-0.02	0.00	0.01	+	†
WEB	K1,3	0.04	-0.01	-0.01	0.00	0.02	+	+
STRANDS	K1,3	0.03	0.01	-0.01	0.00	-0.01	ť	†
AMBITIOU	K1,3	†	ţ	0.03	0.06	0.00	t	†
WAGES	K1,5	+	†	0.02	-0.04	ţ	0.01	+
ALIGNMNT	K1,5	+	†	-0.04	-0.13	+	0.01	+
KINDLETR	3,5	+	†	ţ	ţ	0.00	0.01	†
GROWUP	3,5	+	†	+	÷	0.02	-0.01	†
WHENPAST	3,5	+	†	+	†	0.00	0.00	+
WHENTOOK	3,5	+	†	+	t	0.00	0.03	+
GAVEWHAT	3,5	+	t	+	t	-0.01	0.00	+
KNIGHT	3,5	+	†	ŧ	ţ	0.00	0.04	+
AUTHFEEL	3,5	+	t	ť	ţ	0.00	0.00	†
SAMEHANG	3,5	+	t	+	†	-0.01	0.00	+
TVSHOW	3,5	+	†	t	ţ	0.02	-0.01	+
TANZANIA	3,5	Ť	Ť	ť	Ť	0.02	-0.02	†
KIND OFC	3,5	+	†	+	†	0.00	-0.01	+
ROBBER	3,5	+	t	ť	ţ	0.00	-0.01	†
WORDARTH	3,5	+	†	+	ţ	0.00	-0.04	+
THOSEDAY	3,5	+	†	+	†	-0.02	0.08	+
WHYROUND	3,5	+	†	t	ţ	0.00	-0.03	+
JAMMED	3,5	+	†	+	÷	0.01	-0.02	+
LINECLUE	3,5	+	†	+	ţ	0.00	-0.01	+
BOW	3,5	+	†	+	÷	0.01	-0.01	†
SURPRISE	3,5	+	†	+	ţ	-0.01	0.02	+
TRAIN	3,5	+	†	+	÷	0.01	-0.01	+
TEARING	3,5	+	+	+	+	-0.12	0.03	+
FEELSAFE	3,5	+	+	+	+	-0.01	0.02	+
3THINGS	3,5	Ť	Ť	†	Ť	-0.02	0.00	†

Table 5-2.Reading assessment, actual minus predicted proportion correct: School years 1998–99,
1999–2000, 2001–02, 2003–04, and 2006–07

See notes at end of table.

_	Used in							
Item	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
DEHYDRAT	3,5	Ť	†	Ť	†	-0.02	0.01	Ť
DOMESTIC	3,5	ť	†	ť	†	-0.01	0.02	†
LIKECHDR	3,5	Ť	†	Ť	†	-0.01	0.04	Ť
INFLUENT	3,5	ţ	†	Ť	†	-0.04	0.01	†
DIFFROOM	3,5	ţ	†	Ť	†	0.03	-0.04	†
PROBSOLV	3,5	†	†	Ť	†	0.07	-0.03	†
MAJTHEME	3,5	ţ	†	Ť	†	0.04	-0.02	†
RACHEL	3,5	†	†	Ť	†	0.06	-0.02	†
MICROWAV	3,5	†	†	†	†	-0.02	0.01	†
HOWAUTH	3,5	†	†	†	†	0.03	-0.01	Ť
COMPARCC	3,5	†	†	Ť	†	0.00	-0.01	†
HOAX	3,5	†	†	†	†	-0.03	0.03	Ť
IDEA-CC	3,5	†	†	ť	†	-0.03	0.03	†
GUESS	3,5	†	†	ť	†	0.01	0.00	†
TRUECROP	3,5	ť	†	ţ	†	0.01	-0.01	†
?DISMISS	3,5	ť	†	ť	†	-0.01	0.02	†
BESTWAGM	3,5	ť	†	ť	†	0.04	-0.01	†
BACKPACK	K1,3,5	0.03	0.02	0.02	-0.01	0.00	0.02	†
LISTEN	K1,3,5	0.04	0.01	0.01	0.00	0.00	-0.02	†
RIDEBIKE	K1,3,5	0.07	0.03	0.00	-0.01	0.00	0.01	†
SIZES	K1,3,5	0.03	0.03	0.01	0.00	0.00	-0.02	†
THROUGH	K1,3,5	0.06	0.00	0.01	0.02	0.00	-0.03	†
WTLESS	K1,3,5	Ť	†	0.00	-0.02	0.01	0.01	†
MOISTURE	K1,3,5	ť	Ť	-0.01	-0.04	-0.02	0.02	†
CRITCISM	K1,3,5	†	†	0.00	-0.11	-0.03	0.05	†
PREFRNCE	K1,3,5	ť	ť	0.11	0.08	0.00	-0.02	†
OVATIONS	5,8	ť	Ť	ţ	†	†	-0.06	0.02
DEPART	5,8	ť	Ť	t	†	†	-0.03	0.03
SPRING	5,8	ť	Ť	t	†	†	-0.01	0.01
4CORNERS	5,8	ť	Ť	t	†	†	0.00	0.01
WHY LEFT	5,8	ť	†	ţ	†	†	-0.02	0.02
SUPRT-LA	5,8	ť	†	ť	†	†	-0.06	0.02
ON MESA	5,8	†	+	+	†	†	0.03	-0.01
DOUBT1	5,8	†	+	+	†	†	0.01	0.00
DOUBT2	5,8	†	+	+	†	†	-0.01	0.02
MAINPURP	5,8	t	†	+	†	†	-0.12	0.04
THEORY2	5,8	†	+	†	†	†	-0.03	0.04
TONE	5,8	†	†	†	†	†	0.05	-0.01

Table 5-2.Reading assessment, actual minus predicted proportion correct: School years 1998–99,
1999–2000, 2001–02, 2003–04, and 2006–07—Continued

See notes at end of table.

	Used in							
Item	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
HOWFEEL	3,5,8	Ť	Ť	Ť	Ť	0.00	-0.02	0.05
SIM PROB	3,5,8	Ť	Ť	†	Ť	0.04	-0.03	-0.03
WHYNOT	3,5,8	Ť	†	†	Ť	0.04	-0.04	-0.01
HELPPRB	3,5,8	Ť	†	†	Ť	0.02	-0.02	-0.02
HELPUND	3,5,8	Ť	Ť	Ť	Ť	0.01	-0.02	0.01

Table 5-2.Reading assessment, actual minus predicted proportion correct: School years 1998–99,
1999–2000, 2001–02, 2003–04, and 2006–07—Continued

† Not applicable.

NOTE: Positive numbers correspond to actual proportion correct that is higher than predicted by the IRT model, and negative numbers to actual proportion correct that is lower than predicted. Items are in order of increasing difficulty within grade groups.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

	Used in							
Item	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
2+5MARBL	K1,3	0.01	0.01	0.04	0.01	-0.16	†	†
12BY2s	K1,3	0.00	-0.01	-0.03	0.01	0.00	+	+
3+7PENNY	K1,3	0.02	-0.01	-0.01	0.01	-0.01	†	†
51015 25	K1,3	-0.02	-0.01	-0.04	0.04	0.07	†	†
4+4-2	K1,3	-0.01	0.00	0.02	0.01	-0.04	†	†
HOWMANY\$	K1,3	0.04	0.01	0.02	-0.04	0.08	†	†
12-?PEN	K1,3	0.06	0.04	0.02	-0.03	0.03	†	†
HEADSUP	K1,3	0.05	0.00	0.00	-0.03	0.07	†	†
GOALS	K1,3	0.00	0.00	0.00	0.00	0.00	†	†
CUBES10	3,5	ţ	ţ	ţ	ť	0.00	0.00	†
NEXT78	3,5	ţ	ţ	ţ	ť	-0.01	0.01	†
DO ADD4	3,5	Ť	Ť	Ť	†	0.02	-0.03	†
TIME1030	3,5	Ť	Ť	Ť	†	0.00	0.01	†
NUMBER60	3,5	Ť	Ť	Ť	†	0.00	0.00	†
CUBESIDE	3,5	Ť	Ť	Ť	†	-0.01	0.01	†
NEXT120	3,5	Ť	Ť	Ť	†	0.00	0.00	†
CHART_64	3,5	Ť	Ť	Ť	†	0.00	-0.01	†
BOX 700	3,5	Ť	Ť	Ť	†	0.00	0.00	†
COLORSYM	3,5	Ť	Ť	Ť	†	-0.06	0.03	†
A568214K	3,5	Ť	Ť	Ť	†	-0.01	0.02	†
CARDS579	3,5	Ť	Ť	Ť	†	-0.04	0.02	†
LUISA13	3,5	Ť	ţ	Ť	ţ	0.02	-0.01	†
TALL75	3,5	Ť	ţ	Ť	ţ	-0.06	0.03	†
MARBLES	3,5	Ť	ţ	Ť	ţ	-0.01	0.04	†
BANKER	3,5	Ť	ţ	Ť	ţ	-0.01	0.00	†
SAMEFRAC	3,5	Ť	ţ	Ť	ţ	-0.02	0.02	†
A13_79	K1,3,5	-0.02	-0.03	-0.04	0.04	0.00	-0.01	†
COST_10	K1,3,5	0.04	0.02	0.04	-0.03	0.01	-0.02	†
CARS15_5	K1,3,5	0.03	0.02	0.02	-0.01	0.00	-0.02	†
CANDY8_2	K1,3,5	0.02	0.02	0.04	-0.03	0.01	0.01	†
PATRNC3	5,8	Ť	ţ	Ť	ţ	†	-0.01	0.02
GAMEPTC3	5,8	Ť	Ť	Ť	Ť	Ť	-0.03	0.03
HOOPS C3	5,8	Ť	ţ	Ť	ţ	†	-0.01	0.01
BUDGETC	5,8	ţ	ţ	ţ	ţ	†	-0.02	0.03
FRACTION	5,8	ţ	ţ	ţ	ţ	†	-0.02	0.01
AREAC3	5,8	Ť	Ť	Ť	Ť	Ť	-0.02	0.04

Table 5-3.Mathematics assessment, actual minus predicted proportion correct: School years 1998–99,
1999–2000, 2001–02, 2003–04, and 2006–07

See notes at end of table.

Item	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
PIZZA	5,8	Ť	Ť	Ť	†	Ť	0.01	0.00
PRISMVLC	5,8	†	ţ	ţ	†	†	-0.14	0.12
CARPETC3	5,8	†	ţ	ţ	†	ţ	-0.02	0.03
SPOONS	3,5,8	†	ţ	ţ	†	0.01	0.00	-0.01
PAGES78	3,5,8	†	ţ	ţ	†	0.02	-0.01	-0.02
CHARGE_5	3,5,8	†	ţ	ţ	†	-0.05	0.03	0.06
MARIA310	3,5,8	†	ţ	ţ	†	0.05	0.01	-0.04
PAIR_100	3,5,8	†	ţ	ţ	†	-0.03	0.05	0.00
GREW4	3,5,8	Ť	ţ	Ť	Ť	0.02	-0.02	-0.01
MIN_BLOW	3,5,8	†	ţ	ţ	†	0.04	-0.01	-0.03
EDGECUBE	3,5,8	ŧ	Ť	Ť	Ť	-0.01	-0.02	0.03
MARK_DOT	3,5,8	ŧ	ţ	Ť	ŧ	0.03	0.06	-0.04
TILESCOV	3,5,8	ţ	Ť	Ť	ţ	0.00	0.00	0.00

Table 5-3.Mathematics assessment, actual minus predicted proportion correct: School years 1998–99,
1999–2000, 2001–02, 2003–04, and 2006–07—Continued

† Not applicable.

NOTE: Positive numbers correspond to actual proportion correct that is higher than predicted by the IRT model, and negative numbers to actual proportion correct that is lower than predicted. Items are in order of increasing difficulty within grade groups.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

Item	Used in grades	Round 5	Round 6	Round 7
ROUIMM	3,5	0.00	0.00	ţ
RWINGS	3,5	0.02	-0.08	Ť
ROUFRZ	3,5	-0.01	0.00	Ť
ROUTAP	3,5	-0.02	0.02	Ť
ROUJUN	3,5	0.00	-0.01	Ť
ROUBRN	3,5	0.00	-0.01	Ť
RHEART	3,5	0.00	0.01	Ť
ROUJAR	3,5	0.00	-0.01	Ť
ROUSRF	3,5	0.00	0.00	Ť
RDESRT	3,5	0.01	-0.02	Ť
ROUSOL	3,5	-0.04	0.05	Ť
YBEES	3,5	0.01	-0.03	Ť
ROUBLB	3,5	-0.01	0.02	Ť
ROUGRT	3,5	0.00	0.01	Ť
ROUMCE	3,5	0.01	-0.01	ť
ROUFLY	3,5	-0.01	0.01	ť
ROUSHD	3,5	0.04	-0.05	Ť
BPLNT2	3,5	-0.02	0.01	ť
BPLANT	3,5	0.08	-0.02	Ť
BSOIL	3,5	-0.04	0.02	Ť
BMAMML	3,5	0.05	-0.02	Ť
CUTSCAB	5,8	†	0.00	0.00
PYRAMID	5,8	ţ	0.05	-0.03
EARTHQK	5,8	ţ	0.02	-0.06
GRAVMOON	5,8	ţ	0.01	-0.01
THUNDER	5,8	ţ	-0.01	0.04
WATRGRPH	5,8	ţ	-0.02	0.00
BURIED	5,8	†	0.01	-0.01
SEEDGROW	5,8	†	0.00	0.00
FOXRABIT	5,8	Ť	-0.10	0.06

Table 5-4.Science assessment, actual minus predicted proportion correct: School years 2001–02,
2003–04, and 2006–07

See notes at end of table.

Item	Used in grades	Round 5	Round 6	Round 7
SOLUTION	5,8	Ť	-0.01	0.01
PENCLH2O	5,8	Ť	0.00	0.00
ROUERT	3,5,8	0.02	-0.01	-0.13
YTHEMT	3,5,8	0.03	-0.01	-0.07
YMOON	3,5,8	0.03	-0.02	-0.03
ROUMTN	3,5,8	-0.02	0.03	-0.03
BSOUND	3,5,8	-0.07	0.02	0.01
BSLIDE	3,5,8	-0.07	0.03	-0.01

Table 5-4.Science assessment, actual minus predicted proportion correct: School years 2001–02,
2003–04, and 2006–07—Continued

† Not applicable.

NOTE: Positive numbers correspond to actual proportion correct that is higher than predicted by the IRT model, and negative numbers to actual proportion correct that is lower than predicted. Items are in order of increasing difficulty within grade groups. Science was not tested in kindergarten or first grade.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002, spring 2004, and spring 2007.

5.1.3 IRT Calibration and Scoring

The IRT calibration was carried out using the PARSCALE program as described in chapter 3. Of the 219 items in the reading pool, two eighth-grade items were deleted from the item pool because of differential item functioning (DIF) for a population subgroup (see section 4.2.5). One reading item was similarly deleted in fifth grade. The estimation of reading item parameters and child abilities was based on the remaining 216 unique items that appeared in all forms of the reading assessments, with the 81 items common to two or more of the assessment versions serving to anchor the scale. Four of the 216 items were deleted from the final scale scores so that the scale would be more closely aligned with framework specifications, leaving 212 items in the final reading scale. No mathematics items were deleted because of differential functioning in eighth grade (two had been deleted for this reason in earlier rounds). The K-1, third-grade, fifth-grade, and eighth-grade mathematics scale is based on 174 unique mathematics items in all assessment forms, including 49 common to more than one version of the assessment. The science scale is based on 111 unique items in third, fifth, and eighth grades, including 38 shared by more than one round (no science items were deleted due to differential item functioning). For each item, the IRT calibration resulted in a set of three item parameters that define a logistic function associated with the item. The height of the function at any point along an ability range corresponds to the estimated probability of a correct answer on the item for a person at that ability level. The tables in appendix B show the item parameters in ascending order of difficulty (IRT "b" parameter).

Each of the rounds of data collection, kindergarten through eighth grade (plus the bridge sample), was treated as a separate subpopulation with its own ability distribution for the purpose of IRT calibration. This feature of PARSCALE and other Bayesian approaches to IRT provides for an empirically based shrinkage toward subpopulation means for extreme ability estimates, low and high. This shrinkage is particularly important for a longitudinal study, where the focus is on measuring gain and it is important to avoid floor and ceiling effects. See section 3.2 for additional details. Table 5-5 presents unweighted theta (ability) means and standard deviations for the subpopulations of the reading, mathematics, and science calibrations. The theta estimates are standardized to mean = 0.0 with a standard deviation = 1.0 for all rounds combined. Sample weights are not applied to the theta estimates; they are applied to the assessment scores discussed below.

Table 5-5.IRT theta (ability) means and standard deviations by subpopulation, seven data collection
rounds plus bridge sample: School years 1998–99, 1999–2000, 2001–02, 2003–04, and
2006–07

	Reading		Mathem	Mathematics		ence
Round	Mean	SD^1	Mean	SD	Mean	SD
All rounds combined	0.01	1.02	0.01	1.00	0.00	0.96
Round 1 (fall-kindergarten)	-1.30	0.52	-1.17	0.48	Ť	ť
Round 2 (spring-kindergarten)	-0.72	0.50	-0.69	0.46	Ť	ţ
Round 3 (fall-first grade)	-0.49	0.52	-0.44	0.47	Ť	ţ
Round 4 (spring-first grade)	0.12	0.45	0.06	0.42	Ť	ţ
Second grade bridge sample	0.66	0.26	0.50	0.30	Ť	ţ
Round 5 (spring-third grade)	0.79	0.31	0.72	0.39	-0.62	0.67
Round 6 (spring-fifth grade)	1.05	0.29	1.12	0.40	-0.01	0.66
Round 7 (spring-eighth grade)	1.34	0.38	1.48	0.44	0.98	0.84

[†] Not applicable. ¹ Standard deviation.

NOTE: Statistics are unweighted. Science was not tested in kindergarten or first grade.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99, seven data collection rounds, 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07, plus bridge sample 2001–02.

The IRT scale scores, T-scores, and proficiency scores were derived from the IRT item parameters and ability estimates. As described above and in section 4.1.2, the set of three parameters for each item defines a logistic function corresponding to the probability of a correct answer for a test taker with a given ability level. At each time point, the ability estimate for each child was used in combination with the item parameters to generate a probability for each item. These probabilities were summed over all items in the assessments to get a scale score for each domain representing an estimate of the number of items the child would have answered correctly if he or she had taken all 212 reading items, all 174 mathematics items, or all 111 science items. The T-scores in the database are theta estimates transformed to a metric of mean = 50.0, standard deviation = 10.0 *within* each round, using cross-sectional sample weights.

Proficiency scores required an additional IRT calibration step. Section 4.1.4 describes the selection of a hierarchical series of mastery levels in reading, and another series in mathematics, marked by clusters of four items at each level. Ten such levels were defined in reading, and nine in mathematics, based on items from the K-1, third-grade, fifth-grade, and eighth-grade assessments. Children were judged to have passed a level (score = 1) if they answered at least three of the four items correctly, and to have failed if at least two wrong answers were given (score = 0). Children with fewer than three right or

two wrong answers (because they omitted items, or because the items defining a particular level were not included in the assessment forms they received) were not scored for the purpose of IRT calibration. The proportion of omitted responses in all subjects in all rounds was negligible, so nearly all children had pass or fail scores on the proficiency levels whose items were present on the forms administered to them. After the initial PARSCALE estimates of item parameters and abilities were obtained, parameters for the proficiency levels were estimated. Ability levels were held constant, and the proficiency level clusters (scored as right, wrong, or not administered) were treated as items for estimating item parameters. In essence, this resulted in prediction of mastery level proficiency from estimates of ability levels derived from all items administered to each child. Extremely close fits of the logistic functions to the proportion passing the item-response-based cluster scores (1 or 0) were observed for all levels in all rounds, for both reading and mathematics.

No proficiency levels were defined for the science test because the more diverse curriculum content meant that acquisition of knowledge and skills in science could not be assumed to follow a hierarchical pattern.

The parameters for the reading and mathematics proficiency levels are shown in table 5-6. The very high "a" parameters are consistent with the assumption that four-item clusters are more reliable than single items and do a better job of discriminating among ability levels. It would be very difficult for a low-ability child to pass a four-item cluster by guessing; the guessing parameters (c) were all fixed at zero.

The IRT parameters permit calculation of probability of proficiency at each mastery level in the same manner as described above for individual items. These probabilities are included in ECLS-K user files. Applications of the proficiency probability scores in measuring status and gain are discussed in section 5.3. An additional proficiency score, the highest proficiency level mastered at each round, is described in section 4.1.4.1. Tables A35 and A36 in appendix A present weighted subgroup differences with respect to mastery of the level that represents the modal "highest level" score within each round.

Table 5-6.IRT parameters for reading and mathematics proficiency levels, based on items from
kindergarten, first-grade, third-grade, fifth-grade, and eighth-grade assessments: School
years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

Proficiency		Reading		Mathematics		
	а	b	с	а	b	c
Level 1	3.50	-1.57	0.00	3.50	-1.96	0.00
Level 2	3.26	-1.02	0.00	3.29	-1.24	0.00
Level 3	3.07	-0.73	0.00	4.00	-0.74	0.00
Level 4	4.00	-0.20	0.00	3.93	-0.18	0.00
Level 5	3.00	0.18	0.00	4.77	0.40	0.00
Level 6	3.50	0.60	0.00	6.48	0.81	0.00
Level 7	4.00	0.86	0.00	4.68	1.20	0.00
Level 8	3.05	1.10	0.00	6.70	1.60	0.00
Level 9	5.92	1.55	0.00	6.33	1.93	0.00
Level 10	3.52	2.05	0.00	Ť	Ť	Ť

† Not applicable

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

5.2 Evaluating the K-1-3-5-8 Longitudinal Scale

Section 5.1 described the construction of the longitudinal score scales and IRT calibration of parameters. This section addresses the issue of the validity of the score scales as measures of child achievement and growth between fall-kindergarten and spring-eighth grade. The validity issue is examined from several perspectives:

- Do the tests measure the right content?
- Is the difficulty of the tests suitable for children's ability levels?
- Do the scores constitute a cohesive scale suitable for longitudinal measurement?
- What is the relationship of the cognitive test scores to scores in different rounds and different subjects, and to teacher ratings and child self-ratings?
- How do the ECLS-K results compare with findings from other studies?

5.2.1 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 describes the design of the tests and development of test frameworks (see section 2.1.2). Curriculum experts and teachers provided input with respect to cognitive skills that are both typically taught and developmentally important. Test frameworks, based on those defined in NAEP in each subject were developed accordingly, and test items in each set of assessments were selected to conform as closely as possible to framework specifications. Field test item pools and proposed final form item selections were reviewed by experts, and content and presentation of items were modified in response to their recommendations.

Appendix C illustrates a psychometric perspective on appropriateness of test content. For each item, the assessment version(s) in which it appears are noted: K-1 for the assessment package used for fall- and spring-kindergarten and fall- and spring-first grade (rounds 1 through 4), 3 for the third-grade assessment (round 5), 5 for fifth-grade (round 6), and 8 for eighth-grade (round 7). IRT calibration allows the estimation of performance on each item for *all* rounds, even rounds in which the item was not used. In general, the largest gains in weighted estimated proportion correct are observed in rounds in which the items were actually administered. For example, for items used only in the K-1 assessments, the greatest gains tend to occur in rounds 1 through 4, with relatively little gain later on. Conversely, for items that were introduced in the fifth- and eighth-grade forms, IRT estimates show that very little gain would have been observed in these items if they had been presented in the earlier rounds. The common items used to link K-1 with third-grade forms, third with fifth-grade, or fifth with eighth-grade, tend to show gains across a wider range of rounds. (An exception to the general pattern of assessment forms matching gains is found for certain difficult items that were included in a supplementary reading form designed to avoid a possible ceiling effect in first grade. The supplementary form was administered only to first-graders who had performed unusually well on the standard set of K-1 forms. These items were too difficult for the majority of first-graders, and showed little gain until the third- and fifth-grade rounds). The match of assessment forms to estimated performance gains suggests that the content of the tests reflected what children had been learning during the intervening time periods.

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

Chapter 2 describes the development of two-stage adaptive tests in each subject area for kindergarten and first grade, with similar assessments assembled for the third-, fifth-, and eighth-grade rounds. The adaptive tests were designed to maximize reliability per unit of testing time by matching test difficulty to children's ability level, while minimizing frustration or boredom that could occur if children received tests that were much too difficult or much too easy (see section 2.1.1). Separate assessment packages for K-1, third, fifth, and eighth grades focused on items of appropriate difficulty for the grade(s) in which they were administered, while containing enough overlapping items to support the longitudinal scale. Psychometric results indicate that this approach, the combination of grade-appropriate assessment versions plus alternative second-stage forms within grade, was successful in selecting items of appropriate difficulty for the test takers.

Evidence that the tests 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 test data. Chapter 2 discusses the importance of avoiding floor and ceiling effects, that is, tests that are much too hard (floor effect) or much too easy (ceiling effect) for a substantial number of test takers. Floor and ceiling effects preclude accurate measurement of children at the extremes of the ability distribution. This is particularly important in a longitudinal study, where score scales with floor and ceiling effects can attenuate measurement of gain for the lowest and highest achieving children.

Chapter 4 reviews the operating characteristics of the ECLS-K assessment forms, including the percentages of below-chance (floor effect) and near-perfect (ceiling effect) scores (see section 4.2.1 and table 4-1 for reading; section 4.3.1 and table 4-5 for mathematics; and section 4.4.1 and table 4-9 for science). No floor or ceiling effects were found for the reading and mathematics tests in any round, that is, only a negligible number of children scored below-chance or near-perfect scores on the combined routing and second-stage items. The science test had a borderline floor effect, with about 5 percent of children scoring below-chance in fifth and eighth grades, with an overall effect on less than 1 percent of the overall sample.

Appendix B shows the match of the weighted ability distribution for each round to the whole set of items in the assessment versions used in the grade. While each child received only the routing test plus one selected second-stage form in each round, the difficulty of the whole set of items administered in each round (routing items plus *all* second-stage forms) should reflect the ability level of the whole sample

for that round. For each subject, appendix B lists all items administered in all rounds of the assessments, sorted in ascending order of item difficulty (IRT "b" parameter). The assessment forms in which each item appeared are also noted. The columns for each round of data collection show the mean and standard deviation of theta, the IRT ability estimate. The asterisks in the columns represent the range of abilities two standard deviations below and above the mean, which should include 95 percent of the sample. For example, fall-kindergarten (round 1) children in appendix table B-1 have a weighted reading mean of -1.32 and standard deviation of 0.50 in the IRT theta metric. That corresponds to an expected range of ability between -2.34 and -0.30 for 95 percent of test takers. The difficulty of items in the K-1 reading assessment forms includes this range. A few easier items are also present to prevent floor effects for the lowest achievers in fall-kindergarten. Since the K-1 assessment forms were used for the first four rounds, fall-kindergarten through spring-first grade, the range of difficulty of items in the K-1 reading forms had to extend to at least two standard deviations above the round 4 mean, or at least b = 1.02. Several K-1 items have difficulty parameters beyond this point, as a precaution against ceiling effects for the highest achievers in spring-first grade. In each subject area, the difficulty range of the test items administered more than spans the range of two standard deviations below and above the theta mean for the round. The evidence in table B3 is consistent with the findings shown in table 4-9 for fifth-grade and eighth-grade science: the low level second-stage test could have included one or two more of the easiest items suitable for the lowest achieving children.

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

Evidence presented in appendix D supports the validity of the score scales for longitudinal measurement in two ways. Examination of IRT "a" parameters suggests that the item pools within each subject are strongly related to a single underlying factor that is consistent across rounds from fall-kindergarten through spring-eighth grade. The fit statistics in appendix D demonstrate that the IRT model appropriately represents the test data collected in each round. Tables of proportion correct in appendix C provide an additional perspective on the score scales derived from the IRT estimates.

If each test taker had answered *all* of the items in the kindergarten through eighth-grade item pools at *every* round of data collection, it would be possible to measure the cohesiveness of the scale by observing alpha coefficients and item biserials. Of course, it would have been neither reasonable nor practical to administer the whole item pools to everyone at every round.

The IRT "a" parameters provide the same type of insight into the cohesiveness of a set of test items (see section 3.2.1). This parameter represents item discrimination, or the ability of an item to discriminate, or separate, people whose ability level is above or below the calibrated difficulty of the item. In other words, the "a" parameters indicate how strongly each item is related to the underlying construct being measured by the test, with values of 1.0 or above indicating a strong relationship. Values above 1.0 for most of the items in a test constitute evidence that there is a strong underlying factor.

Of the 212 items in the reading scale, only 13 have "a" parameter values less than 1.0, and five of those are picture-vocabulary items. Most of the rest are based on either listening comprehension or understanding conventions of print. Nearly all of the items measuring reading skills, from simple letter recognition and decoding in kindergarten to comprehension of complex reading passages in the later rounds, have "a" parameters above 1.0, with the exception of three difficult decoding words in fifth grade, two of which were deleted from the score scale. Results for mathematics were quite similar, with only five of 176 items having "a" parameters below 1.0; two of these had low discrimination because they were extremely easy. In earlier rounds, a disproportionate number of mathematics items with low "a" parameters were geometry items, which were identified in the field tests as being slightly weaker than the other mathematics categories with respect to cohesiveness of the scale, but were included in the item pool to conform to framework specifications. This was not the case in the current calibration: only one of the five items with "a" parameters below 1.0 was a geometry item. This suggests that, with the addition of the eighth-grade forms to the item pool, the factor underlying the mathematics scale has a stronger geometry component than before. Examination of the reading and mathematics "a" parameters provides evidence that the item pools and resulting score scales are strongly related to an underlying construct that spans the kindergarten through eighth-grade years.

Results for the science assessment are strikingly different, with "a" parameters for nearly half of the items (51 out of a total of 111) falling below 1.0. This is a consequence of the composition of the science item pool, which is a mix of life science, earth science, and physical science topics. Furthermore, the science assessments did not assume a hierarchical structure in the science curriculum comparable to the patterns for reading and mathematics. In other words, it would be possible for children in some schools to master difficult material relating to the life sciences without having been exposed to basic concepts in earth science, or vice versa. That is the reason that proficiency levels within the science assessments were neither hypothesized nor identified. The relatively low "a" parameters for the science items do not necessarily, however, make IRT methodology inappropriate for calibration of the science

scale. In fact, for all except 6 of the 111 items, "a" parameter values were .60 or above. This suggests that, although there may be multiple factors influencing item responses, they are all related to each other.

Section 5.1.2 explains the use of the fit statistics presented in tables 5-2 through 5-4 in evaluating the functioning of common items tying the score scale together across assessment versions. Appendix D presents the same fit statistics for *all* items in the assessments. In each round, proportion correct for all children who answered each test item was compared with the proportion correct predicted by the IRT model for the same children. The extremely small differences between actual and predicted percent correct for virtually all items at all rounds—even the science items—support the idea that the IRT model appropriately represents the test data collected in each round.

Appendix C shows the weighted proportion correct estimated by the IRT procedures for each item at each round, for *all* children tested. The increase in proportion correct over time, and the fact that increases took place at the rounds expected given the content and difficulty of the items, provides further evidence that the IRT results appropriately model achievement growth.

5.2.4 Relationship of the Cognitive Test Scores to Scores in Different Rounds and Different Subjects, and to Teacher Ratings and Child Self-Ratings

Table 5-7 shows correlations of weighted test scores in each round with scores in the same subject in other rounds. Note that, within each domain, correlations are 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 least closely related to rounds that are more distant. For example, the highest correlation (i.e., best predictor) for round 7 mathematics is the round 6 mathematics measure, with a correlation coefficient of .86. Previous mathematics scores are also strongly correlated with round 7 mathematics, but the relationship becomes weaker going back in time. While mathematics ability at kindergarten entry is a good predictor of eighth-grade achievement (correlation = .64), other factors present in the intervening years presumably have an important influence as well. Measures of family and school circumstances that relate to child achievement are provided in the ECLS-K database. Exploration of the role these variables play in predicting later achievement is beyond the scope of this report.

Subject	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Reading							
Round 1	1.00						
Round 2	.79	1.00					
Round 3	.78	.88	1.00				
Round 4	.67	.77	.82	1.00			
Round 5	.61	.67	.70	.75	1.00		
Round 6	.57	.64	.64	.71	.84	1.00	
Round 7	.56	.56	.59	.62	.73	.79	1.00
Mathematics							
Round 1	1.00						
Round 2	.84	1.00					
Round 3	.80	.85	1.00				
Round 4	.71	.77	.81	1.00			
Round 5	.71	.74	.77	.80	1.00		
Round 6	.68	.70	.73	.77	.88	1.00	
Round 7	.64	.68	.69	.72	.84	.86	1.00
Science							
Round 5	†	†	†	†	1.00		
Round 6	+	+	÷	+	.85	1.00	
Round 7	Ť	+	†	+	.78	.81	1.00

Table 5-7.Correlations of IRT theta score across rounds, by subject: School years 1998–99, 1999–2000,
2001–02, 2003–04, and 2006–07

† Not applicable.

NOTE: Table estimates are based on C1_7SC0 panel weight. Science was not tested in kindergarten or first grade.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

Correlations of weighted scores *across* subjects *within* rounds are presented in table 5-8. These statistics are consistent with estimates from numerous studies. The relationship between reading and mathematics achievement tends to be close to .75 at all ages from early childhood through the end of middle school.

A final perspective on construct validity of the assessments is their relationship with concurrent measures within the ECLS-K survey, namely, the teacher ratings and child self-ratings. These are discussed in chapter 7, section 7.2.

Round	Reading x Mathematics	Reading x Science	Mathematics x Science
Round 1	0.78	÷	Ť
Round 2	0.77	Ť	Ť
Round 3	0.74	Ť	Ť
Round 4	0.72	Ť	Ť
Round 5	0.72	0.71	0.72
Round 6	0.74	0.75	0.75
Round 7	0.71	0.75	0.80

Table 5-8.Correlations of IRT theta score across subjects, by round: School years 1998–99, 1999–2000,
2001–02, 2003–04, and 2006–07

† Not applicable.

NOTE: Table estimates are based on C1 7SC0 panel weight. Science was not tested in kindergarten or first grade.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring, 2004, and spring 2007.

5.2.5 Comparison of ECLS-K Results with Findings from Other Studies

An additional way to validate the ECLS-K measures would be to compare the ECLS-K results with findings of similar studies. Ideally, these "similar studies" would have tests that measure the same content as the ECLS-K tests and have similar formats, administration procedures, reliabilities, and scoring methodology. Children would be sampled from the same population as the ECLS-K (children entering kindergarten in the U.S. in fall 1998, with some sample freshening in later rounds), with adequate sample sizes and comparable sampling and weighting procedures. Children would be in the same grades at the same ages as the ECLS-K sample, and the similar studies would have been conducted in the recent past. Definitions of subpopulations to be compared would be the same for the ECLS-K and the comparison studies. If all of these conditions were met, a finding that the ECLS-K results were similar to those of a similar study would support the validity of the ECLS-K cognitive test scores. Conversely, discrepancies between the results would call into question the validity of the findings of one or both studies. Unfortunately, no published studies could be found that replicate the ECLS-K structure closely enough to expect that findings would be consistent.

A key result that would be important to replicate would be estimates of test score gaps between population subgroups. Numerous studies document the existence of score gaps, especially between Black and White children at various ages and in various subjects. A great deal of work has been done on studying correlates of these gaps, and cross-sectional and longitudinal changes in the gaps. While there is general consensus on factors that influence score gaps, there is by no means consensus on the *size* of the gaps (Jencks and Phillips 1998; Rouse, Brooks-Gunn, and McLanahan 2005). In fact, there is no truly reliable estimate of *the* Black-White score gap, for all of the following reasons, and others:

- Comparability depends on exactly *what* is being measured: verbal tests that focus primarily on vocabulary seem to find larger gaps than reading tests with more diversity of content.
- Timeframe is important: in recent decades, such factors as desegregation, trends in class sizes, and increased preschool attendance have tended to reduce the size of Black-White child score gaps in the early years of school. Findings from recent studies may be quite different from those carried out 10 or 20 years ago (e.g., Grissmer, Flanagan, and Williamson 1998).
- Studies of "stereotype threat" show that context and mode of administration may influence performance, especially for Black children (e.g., Steele et al. 1998).
- Many studies are not designed to be nationally representative, but may be based, for example, on children in a certain type of preschool program, or children in a particular city that may not closely resemble the characteristics of the ECLS-K nationally representative sample.

A literature review and in-depth study of test score gaps is well beyond the scope of this report. However, a few similarities and differences with other findings may be noted that may aid in the evaluation of the consistency of the ECLS-K findings with other studies.

Several studies reported Black-White score gaps for children age 5 or 6, or in kindergarten or first grade, of about one standard deviation, based on the Peabody Picture Vocabulary Test. Some of these studies noted that vocabulary gaps for children of this age are typically larger than gaps found in measures of early reading (Rock and Stenner 2005; Jencks 1998; Phillips, Crouse, and Ralph 1998; Phillips, Brooks-Gunn, et al. 1998). The Black-White score gap in the ECLS-K fall-kindergarten reading test, which contained some picture vocabulary items but primarily focused on early literacy, was indeed smaller: about four-tenths of a standard deviation.

A consensus finding of several studies was that Black-White gaps tend to widen after children enter school (Grissmer, Flanagan, and Williamson 1998; Ferguson 1998). This was consistent with ECLS-K results. In the ECLS-K, the Black-White reading (weighted) score gap increased from about 0.40 of a standard deviation (SD) in the first three rounds to 0.52 SD by spring-first grade, about 0.71 SD in rounds 5 and 6, when most children were in third and fifth grade, and 0.87 in round 7. A

similar pattern was found for mathematics, with an initial fall-kindergarten gap of 0.61 standard deviations widening to 0.82 and 0.84 SD in rounds 5 and 6, and then 0.89 in round 7.

The study that is perhaps most comparable with ECLS-K may be the National Assessment of Educational Progress (NAEP) 2007 assessments in reading and mathematics. Both were large-scale samples representing a national population, in the same year and grade. The content specifications for the ECLS-K tests were derived from NAEP frameworks. Similar IRT methodology was used in producing score scales. Table 5-9 shows reading and mathematics weighted score gaps for selected subgroups for the NAEP 2007 eighth-grade assessment and for the ECLS-K round 7, which consisted primarily of eighth-graders. The NAEP subgroup differences in reading and mathematics scores were similar to the differences found in the ECLS-K round for the male/female comparison and for White children compared with Black children. The reading gap for the male/female comparison is larger for NAEP than for the ECLS-K but the same in mathematics. Conversely, the reading and mathematics gaps for the Black and White child comparisons are larger for the ECLS-K than for NAEP. Statistics for White/Hispanic score gaps are included in table 5-9 although race/ethnicity for Hispanic children is defined differently in NAEP and ECLS-K.

Subgroup gaps in standard deviation units	NAEP Grade 8	ECLS-K Round 7
Reading		
Female - male	.29	.20
White - Black	.77	.87
White - Hispanic	.71	Ť
White - Hispanic, race specified	Ť	.43
White - Hispanic, race not specified	ť	.71
Mathematics		
Female - male	06	06
White - Black	.86	.89
White - Hispanic	.72	Ť
White - Hispanic, race specified	Ť	.33
White - Hispanic, race not specified	ť	.52

Table 5-9.	Subgroup gaps in standa	rd deviation units, NAEP	and ECLS-K: School	year 2006-07
				2

† Not applicable.

NOTE: Table estimates for the ECLS-K are based on C1_7SC0 panel weight.

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2007 Reading and Mathematics Assessments, and Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007. As similar as the NAEP and ECLS-K assessments are in many respects, there are also some important differences that relate to the comparability of measurements of gaps:

- The NAEP used a cross-sectional sample of children in eighth grade in 2007; the ECLS-K sample was a longitudinal follow-up of a kindergarten sample. Most of the children tested in ECLS-K round 7 in 2007 were in eighth grade, but about 11 percent of children were not in the modal grade.
- The NAEP cross-sectional sample could be expected to contain more recent immigrants than the ECLS-K longitudinal sample. The ECLS-K round 7 children tested in spring 2007 had all joined the sample in kindergarten or first grade, during the 1998–99 or 1999–2000 school year, so they had been attending school in the U.S. for at least 7 years. The NAEP sample consisted of children in eighth grade in 2007 and included children whose early schooling may have taken place in another country with instruction in a language other than English.
- The NAEP had two different sources for race variables: school records and child self-report (table 5-9 shows race/ethnicity from school records). The ECLS-K used a composite race/ethnicity variable, derived from parent interviews in most cases, and from a variety of other sources when parent reports were unavailable.
- The NAEP reported scores for Hispanics as a group, while the ECLS-K had separate categories for Hispanic, race specified and Hispanic, race not specified. Table 5-9 shows larger White-Hispanic score gaps for NAEP than for ECLS-K. This is probably due at least in part to the presence of recent immigrants in the NAEP sample but not in the ECLS-K longitudinal group as noted above.

5.3 Applications

This section describes issues in selection and use of scores for analyzing status and gain in cognitive skills. Appendix A includes weighted score breakdowns by gender, ethnicity, socioeconomic status (SES), and school type for all of the eighth-grade direct cognitive measures. For measures that can be compared with the analogous scores in earlier rounds, results for rounds 1 through 6 are included in the tables as well. Examination of similarities and differences, within and across rounds, may suggest research questions that can be addressed by the ECLS-K data and assist with formulation of analysis models. Breakdowns for earlier rounds' reading and science cluster scores are not included in appendix A because these scores do not exist in round 7. Documentation of performance on these scores can be found in appendix A, tables A13 through A19, of the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Psychometric Report for the Fifth Grade* (NCES 2006–036rev) (Pollack, Atkins-Burnett et al. 2005).

5.3.1 Choosing Appropriate Scores for Analysis

Each of the types of scores described earlier measures children's achievement from a slightly different perspective. The choice of the most appropriate score for analysis purposes should be driven by the context in which it is to be used:

- a measure of overall achievement versus achievement in specific skills;
- an indicator of status at a single point in time versus growth over time; and
- a criterion-referenced versus norm-referenced interpretation.

5.3.1.1 Item Response Theory-Based Scores

The scores derived from the IRT model (IRT scale scores, T-scores, proficiency probabilities) are based on all of the child's responses to a subject area assessment. That is, the pattern of right and wrong answers, as well as the characteristics of the assessment items themselves, are used to estimate a point on an ability continuum. This ability estimate, theta, then provides the basis for criterion-referenced and norm-referenced scores.

The IRT scale scores are overall, criterion-referenced, measures of status at a point in time. They are useful in identifying cross-sectional differences among subgroups in overall achievement level and provide a summary measure of achievement useful for correlational analysis with status variables, such as demographic, school type, or behavioral measures. The IRT scale scores may be used as longitudinal measures of overall growth. However, gains made at different points on the scale have qualitatively different interpretations. For example, children who make gains in recognizing letters and letter sounds are learning very different lessons from those who are making the jump from reading words to reading sentences, although the gains in number of scale score points may be the same. Comparison of gain in scale score points is most meaningful for groups that started with similar initial status.

The standardized scores (T-scores) are also overall measures of status at a point in time, but they are norm-referenced rather than criterion-referenced. They do not answer the question, "What skills do children have?" but rather, "How do they compare with their peers?" The transformation to a familiar metric with a mean of 50 and standard deviation of 10 facilitates comparisons in standard deviation units. T-score means may be used longitudinally to illustrate the increase or decrease in gaps in achievement among subgroups over time. T-scores are not recommended for measuring individual gains over time. The IRT scale scores or proficiency probability scores are more suitable or appropriate for that purpose.

Proficiency probability scores, derived from the overall IRT model, are criterion-referenced measures of proficiency in specific skills. Because each proficiency score targets a particular set of skills, these scores are ideal for studying the details of achievement, rather than the single summary measure provided by the IRT scale scores and T-scores. They are useful as longitudinal measures of change because they show not only the extent of gains but also where on the achievement scale the gains are taking place. Thus, they can provide information on differences in skills being learned by different groups, as well as the relationships of skill gains with processes, both in and out of school, that correlate with learning specific skills. For example, high-SES kindergarten children showed very little gain in the lowest reading proficiency level, letter recognition, because they were already proficient in this skill at kindergarten entry. At the same time, low-SES children made big gains in basic skills, but most had not yet made major gains in reading words and sentences by the end of kindergarten. Similarly, the best readers in eighth grade may be working on learning to make evaluative judgments based on biographical reading material, which would show up as large gains in reading level 9. Less skilled readers may show their largest gains between fifth and eighth grade at levels 6 or 7, literal inference and extrapolation. The proficiency level at which the largest change is taking place is likely to be different for children with different initial status, background, and school setting. Changes in proficiency probabilities over time may be used to identify the process variables that are effective in promoting achievement gains in specific skills.

5.3.1.2 Scores Based on Number Right for Subsets of Items (Non-IRT Based Scores)

The **routing test number-right** and **item cluster scores** do not depend on the assumptions of the IRT model. They are derived from item responses on specific subsets of assessment items, rather than estimates based on patterns of overall performance. Highest proficiency level mastered also, in theory, is derived from item responses, although IRT-based estimates were used to supplement item response data where necessary.

Routing test number-right scores for the eighth-grade reading, math, and science assessments are based on the 10 items administered in each subject (25, 18, and 21 items, respectively, in fifth grade; 15, 17, and 15 items for the same subjects in third grade; and 20, 16, and 12 items for the K-1

reading, math, and general knowledge assessments). They target specific sets of skills and cover a broad range of difficulty. These scores may be of interest to researchers because they are based on a specific set of assessment items, which was the same for all children who took the eighth-grade assessment. Note that comparisons of routing test number-right scores may be made *within* rounds 1 through 4, because the same set of assessment forms was used in those rounds, and all children received the same sets of routing items. However, scores on the third-, fifth-, and eighth-grade routing tests were each based on different and more difficult sets of items. The eighth-grade routing test number-right scores should *not* be compared with the routing test number-right scores for earlier rounds. Appendix A presents breakdowns for routing test number-right scores only for the eighth-grade forms; statistics for the routing scores for earlier grades can be found in the psychometric reports for each round.

Item cluster scores in reading (e.g., Decoding Score Gr 5) and science (e.g., Life Science Gr 5) are based on a count of the number correct for a particular set of items. Users may wish to relate these scores to process variables to get a perspective that is somewhat different from that of the hierarchical levels of skills. However, with only three to seven items in each of these item cluster scores, reliabilities tend to be relatively low (see sections 4.2.3 and 4.4.3).

Highest proficiency level mastered is based on the same sets of items as the proficiency probability scores but consists of a set of dichotomous pass/fail scores, reported as a single highest mastery level. Pass/fail on each of the individual levels in the set is based on whether children were able to answer correctly at least three out of four actual items in each cluster. Over all rounds of data collection, for about 33 percent of these scores in reading, and about 20 percent in mathematics, the item data were supplemented with IRT-based estimates to avoid complications associated with non-random missing data. The highest proficiency level mastered should be treated as an ordinal variable.

5.3.1.3 Choosing the Correct Sample Weight

The ECLS-K database contains several versions of sample weights, designed to identify children participating in selected rounds and produce national estimates accordingly. Cross-sectional weights should be used only when analyzing data from a single round of data collection. When multiple rounds are involved, as in predicting outcomes in later rounds from variables measured earlier, a panel weight is appropriate. Panel weights are defined for specific combinations of rounds. If analysis of round 7 outcomes depends on inputs from *all six* previous rounds, the C1_7SCO panel weight can be selected.

This panel weight has a value of zero for any child who did not participate in one or more rounds. It is important to remember that the round 3 (fall-first grade) data collection was based on a small subsample of approximately 30 percent of the longitudinal sample. Selecting the C1_7SCO panel weight will, in effect, delete all cases from the analysis that were not part of the fall-first grade subsample. While weighted estimates may not be affected very much, significance tests depend on *un*weighted sample sizes, so findings of statistical significance, especially for analysis of population subgroups, could be severely affected. If fall-first grade variables are not specifically required, using the C1_7FCO panel weight, which depends on participation in rounds 1, 2, 4, 5, 6, and 7, but not round 3, would increase sample sizes substantially. Additional details on selection and application of sample weights can be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks (NCES 2009–004) (Tourangeau, Nord et al. forthcoming).*

5.3.2 Notes on Measuring Gains

This section outlines approaches to measuring gains that rely on multiple criterionreferenced points to identify different patterns of child growth. It describes how analysts might use the proficiency probability scores to address policy questions dealing with subgroup differences in achievement growth over time.

Traditional approaches using a total scale score to measure change may yield uninformative if not misleading results. For example, analysis of the gain in weighted total scale score points in reading between fall- and spring-kindergarten shows an average increase of about 11 points, about 22 points between spring-third grade and spring-fifth grade, and about 20 points between spring fifth-grade and spring eighth-grade. Subgroup analysis shows nearly identical average gains of about the same magnitude for groups broken down by sex, race/ethnicity, SES, and school type, even though the *mean scores* for the subgroups are quite different. Similarly, each of these groups gained about 10 points, on average, on the mathematics scale during kindergarten, and about 23 points between third grade and fifth grade, and about 18 points between fifth and eighth grade, again starting from very different initial statuses. (The similarity in scale score gains between reading and mathematics is coincidental; there is no claim that the same score or amount of gain in different subjects represents a comparable level of achievement or gain.)

It would be incorrect to conclude that, because different subgroups of children are gaining quantitatively the same number of scale score points, they are learning the same things, or that these gains are qualitatively comparable in any sense. The problem is nonequivalence of scale units: children who gain 10 points at the low end of the scale during kindergarten, for example, by mastering letter recognition and letter sounds, are not learning the same things as more advanced children in the same grade who are achieving their 10-point gains by learning to read words and sentences. Nor can gains in comprehension of reading passages in the later rounds be considered equivalent to gains of the same number of points in basic skills in the early elementary years.

The use of adaptive assessments increases the reliability of individual assessment scores by removing the sources of floor and ceiling effects. When assessment forms are matched to children's ability levels, all children have an equal chance to gain on the vertical scale. Depending on how adaptive the measure is, how the scale is constructed, and how even-handed the educational treatment, one may not observe large differences among individual children's amounts of gain in total scale score points. Individual and group differences in the *amount* of gain given a fairly standard treatment (e.g., a year or two of schooling) can be relatively trivial compared with individual and group differences in *where* the gains take place. It is more likely that one will see substantial subgroup differences in initial status than in scale score point gains, suggesting that the gains being made by individuals at different points on the score scale are qualitatively different. Thus, analysis of the total IRT scale score without explicitly taking into consideration where the gain takes place tells only part of the story.

The ECLS-K design utilized adaptive assessments to maximize the accuracy of measurement and minimize floor and ceiling effects and then to develop an IRT-based vertical scale with multiple criterion-referenced points along that scale. These points, the ten reading and nine mathematics proficiency levels that were described in chapter 4, model critical stages in the development of skills. Criterion-referenced points serve two purposes at the individual level: (1) they provide information about changes in each child's mastery or proficiency at *each* level and (2) they provide information about *where* on the scale the child's gain is taking place. This provides analysts with two options for analyzing achievement gains and relating them to background and process variables. First, gains in probability of proficiency at any level may be aggregated by subgroup and/or correlated with other variables. Second, the location of maximum gain may be identified for each child by comparing the gains in probability for all of the levels and focusing on the skills the child is acquiring during a particular time interval. The probabilities of proficiency at any level may be averaged to estimate the proportion of children mastering the skills marked by that level. For example, the weighted spring-first grade mean for mathematics level 5, "Multiply/Divide," was 0.22, analogous to 22 percent of the first-grade population demonstrating mastery of this set of items. The mean probability at the end of third grade, 0.75, is equivalent to a population mastery rate of 75 percent (see table A30). While most children were making their largest gains between first and third grade at level 5, a small number of children were advancing their skills in solving word problems based on rate and measurement, level 7. The mastery rate for level 7 advanced from near zero at the end of first grade to 13 percent at the end of third grade (shown in table A32). The table breakdowns demonstrate that these proportions and the average gains in the proportions for this particular skill are quite different for subgroups of children defined by various demographic and school-process categories. Similarly, gains at each level between any selected round and a subsequent round may be computed for individual children and treated as outcome variables in multivariate models that include background and process measures.

Another approach to the analysis of gain entails computing differences in probabilities of proficiency between any two rounds for *all* of the proficiency levels. The largest difference marks the mastery level where the largest gain for a given child is taking place: the "locus of maximum gain." The locus of maximum gain is likely to vary for different subgroups of children categorized according to variables of interest. Once having identified mutually exclusive groups of children according to the proximity of their gains to each of the critical points on the developmental scale, one can treat the different types of gains as qualitatively different outcome measures to be explained by background and process variables.

Each different analytical approach provides a different perspective with respect to understanding child growth. While comparisons of scale score means may be used to capture information about children at a single point in time, analysis of gains in probability of proficiency is more likely to provide useful information about the contribution of background and process variables to gains in achievement over time. Examples of these approaches can be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05) (Rock and Pollack 2002).

Another important issue to be considered in analyzing achievement scores and gains is assessment timing: children's age at first assessment, assessment dates, and the time interval between successive assessments. Assessment dates ranged from September to November for fall-kindergarten and fall-first grade data collections, and from March to June for spring rounds. At kindergarten entry, boys, on average, tend to be older than girls. Children assessed in November of their kindergarten year may be expected to have an advantage over children assessed in the first days or weeks of school. Substantial differences in intervals between assessments may also affect analysis of gain scores. Children assessed in November and June of kindergarten or first grade have more time to learn skills than children assessed in November and March. These differences in intervals may have a relatively small effect on analysis results for long time intervals, such as measuring gains from spring-first grade to spring-third grade, but may be more important within grade, especially fall-to-spring kindergarten. In designing an analysis plan, it is important to consider whether and how differences in ages, assessment dates, and intervals may affect the results, to look at relationships between these factors and other variables of interest, and to compensate for differences if necessary. More details can be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05) (Rock and Pollack 2002).

In third grade and later rounds, about 10 percent of ECLS-K participants were not in the modal grade for the sample; most of these children were one grade level behind. It is important to keep in mind that, although documentation refers to the "fifth-grade" or "eighth-grade" round, the ECLS-K consists of a follow-up of kindergarten entrants, not a representative sample of children in the modal grade. Researchers will have to consider the possible implications of this situation in the design and interpretation of analyses based on the data collected.

6. PSYCHOMETRIC CHARACTERISTICS OF THE SOCIOEMOTIONAL MEASURES

Chapter 6 describes the development of the socioemotional measures in the student questionnaire, which included two sets of scales to measure the socioemotional development of children. The first was the self-description questionnaire, which was used to determine how children thought and felt about themselves both academically and socially. The second set of scales consisted of the Self-Concept and Locus of Control scales adapted from the student questionnaire used in the National Education Longitudinal Study of 1988 (NELS:88). The Self-Concept scale was derived from the Rosenberg Self-Esteem Scale (RSE) (Rosenberg 1965). These scales asked children about their perceptions about themselves and the amount of control they had over their own lives. This chapter also provides details of the psychometric characteristics of these scales. Weighted means of the socioemotional measures are presented to provide estimates of child-level assessment scores for eighth grade that are representative of the population cohort. Estimates of the scales of the psychometric characteristics of these data (rather than the population).

6.1 Self-Description Questionnaire

Beginning in the third-grade data collection in the ECLS-K, children were asked to provide self-assessments of their academic and social skills in the self-description questionnaire. For the eighth-grade self-description questionnaire, children rated their perceived competence and interest in reading and mathematics. Children also rated self-perceived problems or sources of problems on the Internalizing Problems scale, which included items on sadness, loneliness, and anxiety. Items for the reading and mathematics scales were drawn from the Self Description Questionnaire (SDQ) II,¹ which was designed for children in middle school. Items for the eighth-grade Internalizing Problems scale,² as recommended by the Content Review Panel because these items better reflected the constructs that the study intended to measure and also allowed for comparison with previous rounds of data collection. For further description of the self-description questionnaire, see chapter 2, section 2.2.

¹ The items were adapted with permission from the Self Description Questionnaire (SDQ) II), from *Self Description Questionnaire (SDQ) II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept. An interim test manual and a research monograph, by H.W. Marsh (Sydney: University of Western Sydney, SELF Research Centre, 1992). (Original work published in 1990)*

² The Internalizing Problems scale was developed for the ECLS-K study.

In the three scales of the eighth-grade self-description questionnaire, children rated whether each item was "not at all true (1)," "a little bit true (2)," "mostly true (3)," or "very true (4)." The scores on all three scales represent the mean rating of the items included in the scales. Children who responded to the eighth-grade self-description questionnaire answered virtually all of the questions,³ so treatment of missing data was not an issue. As with most measures of socioemotional behaviors, the distributions on these scales are skewed (negatively skewed for the positive social behavior scales and positively skewed for the problem behavior scales).

6.1.1 Reliability Analysis

Table 6-1 presents the internal consistency reliability estimates of the eighth-grade selfdescription questionnaire scales, as measured by Cronbach's coefficient alpha. The Cronbach's coefficient alpha for the Perceived Interest and Competence in Math is similar to that found by the scale's authors (alpha = .89) (Ellis, Marsh, and Richards 2002). However, the coefficient for the eighth-grade Perceived Interest and Competence in Reading scale is lower than that found by the scale's authors (alpha = .88) (Ellis, Marsh, and Richards 2002). The coefficient alpha for the eighth-grade Internalizing Problem Behaviors scale is consistent with the findings from the ECLS-K fifth-grade data (alpha = .79) (Pollack, Atkins-Burnett et al. 2005).

Table 6-1.	Reliability estimates for scores of the self-description questionnaire scales, spring-eighth
	grade: School year 2006–07

			Published alpha coefficient of
Description	Number of items	Alpha coefficient	SDQ
Perceived Interest/Competence — Reading	4	.76	.88
Perceived Interest/Competence — Math	4	.89	.89

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

 $^{^{3}}$ There were very few nonresponse values in the self-description questionnaire item data. By case, children had a nonresponse value on an average of 0.18 items on the self-description questionnaire measure. By item, the average percent of cases that had a nonresponse value on the self-description questionnaire items was approximately 1 percent.

6.1.2 Factor Analysis

To further explore the stability of the self-description questionnaire scales, principal components factor analyses were conducted for the all items. The factor analyses with the self-description questionnaire items specified the extraction of three factors. The eigenvalues and proportion of variance accounted for by each component are listed in table 6-2. These three factors account for a total of 52.6 percent of the variance.

Table 6-2.Eigenvalues and proportion of variance accounted for by the three factors extracted in
principal components factor analysis with self-description questionnaire data: School year
2006–07.

Characteristic	Factor 1	Factor 2	Factor 3
Eigenvalue	3.17	2.98	2.21
Proportion of variance accounted for by component	19.8	18.6	13.8

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

6.1.3 Mean Scores

Table 6-3 presents the weighted means and standard deviations for the self-description questionnaire subscales.

Table 6-3.Self-description questionnaire weighted means and standard deviations, spring-eighth grade:
School year 2006–07

Description	Weighted mean	Standard deviation
Perceived Interest/Competence — Reading	2.52	.78
Perceived Interest/Competence — Math	2.62	.91
Internalizing Problems	2.03	.57

NOTE: Table estimates based on C7CW0 weight. The range of values is 1-4.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

Self-description questionnaire subscale score statistics for subpopulations are presented in tables 6-4 through 6-6. Children who had been retained (sixth- or seventh-graders in this round) rated themselves lower in the academic interest/competence areas and as having more internalizing behavior problems. Please note that the different sample totals (Ns) for the different scales in tables 6-4 through 6-6

are due to the missing data patterns for these sets of measures. A small number of children did not respond to all of the self-description questionnaire items and thus have scores on some, but not all, of the self-description questionnaire scales.

Table 6-4.	Score breakdown, self-description questionnaire, Perceived Interest/Competence in Reading,
	by eighth-graders, sixth- and seventh-graders, and population subgroup: School year
	2006–07

	Eighth-graders		Sixth- and	Sixth- and seventh-graders		
Characteristic	Number	Mean	SD^1	Number	Mean	SD
Total sample	8,324	2.55	0.75	888	2.36	0.73
Sex						
Male	4,076	2.39	0.73	554	2.26	0.73
Female	4,248	2.71	0.74	334	2.52	0.70
Race/ethnicity						
White, non-Hispanic	5,230	2.58	0.72	441	2.30	0.70
Black, non-Hispanic	732	2.60	1.01	200	2.45	0.89
Hispanic, race specified	697	2.47	0.79	81	2.34	0.64
Hispanic, race not specified	741	2.34	0.75	82	2.28	0.63
Asian	483	2.71	0.54	24	2.64	0.62
Hawaiian, other Pacific						
Islander	104	2.14	0.51	2	‡	†
American Indian/Alaska					•	,
Native	134	2.31	0.55	44	2.33	0.47
More than one race,						
non-Hispanic	195	2.61	0.62	13	2.33	0.62
Socioeconomic status						
First quintile (lowest)	1,021	2.34	0.76	252	2.30	0.72
Second quintile	1,388	2.40	0.74	189	2.37	0.77
Third quintile	1,556	2.56	0.76	130	2.31	0.66
Fourth quintile	1,627	2.61	0.74	93	2.30	0.73
Fifth quintile (highest)	1,989	2.77	0.68	69	2.64	0.54
School type						
Public school	6,801	2.54	0.78	795	2.34	0.74
Private school	1,482	2.64	0.56	78	2.60	0.51

† Not applicable.

¹ Reporting standards not met. ¹ Standard deviation.

NOTE: Table estimates are based on C7CW0 weight. The range of possible values is 1 to 4. Subgroup counts do not sum to total because demographic variables are missing for some cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2007.

	Eightl	n-graders		Sixth- and	seventh-g	graders
Characteristic	Number	Mean	SD^1	Number	Mean	SD
Total sample	8,323	2.63	0.90	894	2.52	0.93
Sex						
Male	4,075	2.69	0.89	560	2.58	0.88
Female	4,248	2.58	0.90	334	2.40	0.99
Race/ethnicity						
White, non-Hispanic	5,230	2.65	0.85	442	2.48	0.87
Black, non-Hispanic	732	2.65	1.24	203	2.50	1.18
Hispanic, race specified	697	2.53	0.97	81	2.57	0.76
Hispanic, race not specified	741	2.51	0.96	83	2.46	0.82
Asian	482	2.85	0.38	24	2.37	0.62
Hawaiian, other Pacific						
Islander	104	2.60	0.68	2	‡	t
American Indian/Alaska					-	
Native	134	2.57	0.78	44	2.86	0.70
More than one race, non-						
Hispanic	195	2.68	0.86	14	3.06	0.84
Socioeconomic status						
First quintile (lowest)	1,020	2.59	1.07	254	2.52	0.96
Second quintile	1,388	2.57	0.93	190	2.49	0.97
Third quintile	1,556	2.60	0.91	130	2.55	0.83
Fourth quintile	1,627	2.64	0.87	94	2.45	0.83
Fifth quintile (highest)	1,989	2.79	0.76	70	2.92	0.76
School type						
Public school	6,800	2.64	0.94	801	2.51	0.94
Private school	1,482	2.60	0.68	78	2.51	0.74

Table 6-5.Score breakdown, self-description questionnaire, Perceived Interest/Competence in
Mathematics, by eighth-graders, sixth- and seventh-graders, and population subgroup:
School year 2006–07

† Not applicable.

‡ Reporting standards not met.

¹ Standard deviation.

NOTE: Table estimates are based on C7CW0 weight. The range of possible values is 1 to 4. Subgroup counts do not sum to total because demographic variables are missing for some cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.
	Eigl	nth-graders		Sixth- and	seventh-g	raders
Characteristic	Number	Mean	SD^1	Number	Mean	SD
Total sample	8,324	2.02	0.54	891	2.12	0.57
Sex						
Male	4,076	1.95	0.54	556	2.05	0.56
Female	4,248	2.09	0.54	335	2.23	0.57
Race/ethnicity						
White, non-Hispanic	5,230	2.01	0.51	441	2.03	0.51
Black, non-Hispanic	732	1.96	0.72	201	2.20	0.71
Hispanic, race specified	697	2.10	0.61	82	2.32	0.58
Hispanic, race not specified	742	2.11	0.59	82	2.18	0.54
Asian	482	2.10	0.38	24	2.05	0.44
Hawaiian, other Pacific						
Islander	104	2.08	0.40	2	‡	ť
American Indian/Alaska					•	
Native	134	2.02	0.49	44	1.95	0.44
More than one race, non-						
Hispanic	195	1.93	0.52	14	1.91	0.39
Socioeconomic status						
First quintile (lowest)	1,020	2.15	0.64	254	2.14	0.59
Second quintile	1,388	2.02	0.55	189	2.07	0.59
Third quintile	1,556	2.01	0.54	130	2.09	0.49
Fourth quintile	1,627	1.99	0.52	93	2.06	0.61
Fifth quintile (highest)	1,989	1.96	0.45	70	2.04	0.48
School type						
Public school	6,800	2.03	0.57	798	2.12	0.58
Private school	1,483	1.96	0.39	78	2.00	0.50

Table 6-6. Score breakdown, self-description questionnaire, Internalizing Problems, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006-07

† Not applicable.

¹ Reporting standards not met. ¹ Standard deviation.

NOTE: Table estimates are based on C7CW0 weight. The range of possible values is 1 to 4. Subgroup counts do not sum to total because demographic variables are missing for some cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2007.

Intercorrelations with other scales are presented in table 7-12 in chapter 7.

6.2 Self-Concept and Locus of Control Scale Scores

As noted earlier, the Self-Concept and Locus of Control scales were adopted from the National Education Longitudinal Study of 1988 (NELS:88). These scales asked children about their self-perceptions and the amount of control they had over their own lives. Items were drawn from the NELS:88 student questionnaire and asked children to indicate the degree to which they agreed with 13 statements about themselves. Statements reflected perceptions children might have about themselves and about how much control they felt they had over their own lives. Children rated whether they "strongly agree," "agree," "disagree," or "strongly disagree" with each item.

In order to be as comparable as possible with NELS:88, scale scores were calculated with the same procedures as NELS:88. Some items were positively worded, and some were negatively worded. As a result, scoring for some items was reversed to provide an appropriate score. For the Self-Concept scale, three of the seven items in the scale were reverse scored before performing computations, so that higher scores indicate more positive self-concept (see table 6-9 for items that were reversed scored). The seven items in the scale were then standardized separately to a mean of zero and a standard deviation of 1. The scale score is an average of the seven standardized scores.

For the Locus of Control scale, five items were reverse scored so that higher scores indicate greater perception of control over one's own life:

- I don't have enough control over the direction my life is taking.
- In my life, good luck is more important than hard wok for success.
- Every time I try to get ahead, something or somebody stops me.
- My plans hardly every work out, so planning only makes me unhappy.
- Chance and luck are very important for what happens in my life.

The six items in the scale were then standardized separately to a mean of zero and a standard deviation of 1. The scale score is an average of the six standardized scores.

Children who responded to the Self-Concept and Locus of Control items answered virtually all of the questions,⁴ so treatment of missing data was not an issue.

6.2.1 Reliability Analysis

Table 6-7 presents the internal consistency reliability estimates of the Self-Concept and Locus of Control scales, as measured by Cronbach's coefficient alpha. The coefficient alpha for both scales are consistent with the findings from the NELS:88 data ($alpha_{Self-Concept} = .79$, $alpha_{Locus of Control} = .68$) (Ingels et al. 1990).

Table 6-7.Self-Concept and the Locus of Control scale reliabilities (alpha coefficient): School year2006–07

Description	Number of items	Alpha coefficient
Self Concept	7	.81
Locus of Control	8	.75

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

6.2.2 Factor Analysis

To further explore the stability of the Self-Concept and Locus of Control scales, principal components factor analyses were conducted for these items from the student questionnaire. The factor analyses with the Self-Concept and Locus of Control items specified the extraction of two factors. The eigenvalues and proportion of variance accounted for by each component are listed in table 6-8. These two factors account for a total of 46.7 percent of the variance.

⁴ There were very few non-response values in the Self-Concept and Locus of Control item data. By case, children had a non-response value on an average of 0.10 items on the Self-Concept and 0.08 items on the Locus of Control scales. By item, the average percent of cases that had a non-response value on these scale items was 1.5 percent.

Table 6-8.Eigenvalues and proportion of variance accounted for by the two factors extracted in
principal components factor analysis with Spring 2007 Self-Concept and Locus of Control
data: School year 2006–07

Characteristic	Factor 1	Factor 2
Eigenvalue	4.59	1.47
Proportion of variance accounted for by component	35.3	11.3

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

The Varimax rotated factor pattern was used to examine the factor structure for each of the two factors (see table 6-9). Variables with factor loadings of $r_{xc} > +0.5$ or $r_{xc} < -0.5$ were identified as loading on the factor. The resultant factor patterns of the two factors matched the structure of the Self-Concept (factor 1) and Locus of Control (factor 2) scales except for one variable: *When I make plans, I am almost certain I can make them work*. This variable was originally mapped onto the Locus of Control scale, but the results of the factor analyses shows this variable loading onto the Self-Concept scale. The analyses of the Cronbach's coefficient alpha of the Locus of Control scale indicate that this item does have a low correlation with the total score (r = .26); nevertheless, dropping this item would not result in a notable increase in the alpha coefficient.

Table 6-9.Varimax rotated factor patterns for the two factors extracted in principal components factor
analysis with Self-Concept and Locus of Control item data: School year 2006–07

Item	Factor 1	Factor 2
I feel good about myself.	.75	.06
I don't have enough control over the direction my life is taking (reverse coded).	.30	.56
In my life, good luck is more important than hard work for success (reverse coded).	04	.69
I feel I am a person of worth, the equal of other people.	.64	.09
I am able to do things as well as most other people.	.61	.10
Every time I try to get ahead, something or somebody stops me (reverse coded).	.27	.57
My plans hardly ever work out, so planning only makes me unhappy (reverse coded).	.42	.54
On the whole, I am satisfied with myself.	.72	.18
I certainly feel useless at times (reverse coded).	.50	.50
At times I think I am no good at all (reverse coded).	.52	.50
When I make plans, I am almost certain I can make them work.	.58	.09
I feel I do not have much to be proud of (reverse coded).	.51	.50
Chance and luck are very important for what happens in my life (reverse coded).	14	.71

NOTE: Bold type indicates on to what factor (1 or 2) the variable more strongly loaded.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

6.2.3 Mean Scores

Table 6-10 presents the weighted means and standard deviations for the Self-Concept and Locus of Control scales.

Table 6-10.Self-Concept and Locus of Control weighted means and standard deviations, spring-eighth
grade: School year 2006–07

Description	Weighted mean	Standard deviation
Self Concept	0.00	.70
Locus of Control	0.02	.64

NOTE: Items were standardized to a mean of 0 and a standard deviation of 1. Scale scores are averages of the respective standardized item scores. Table estimates are based on C7CW0 weight.

Self-Concept and Locus of Control scale score statistics for subpopulations are presented in tables 6-11through 6-12. Generally, children who had been retained (sixth- or seventh-graders in this round) rated themselves lower in terms of self-concept and as having less control over their own lives.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

	Eigl	nth-graders		Sixth- and	d seventh-g	raders
Characteristic	Number	Mean	SD^1	Number	Mean	SD
Total sample	8,314	0.01	0.69	890	-0.16	0.69
Sex						
Male	4,072	0.06	0.66	558	-0.16	0.69
Female	4,242	-0.03	0.71	332	-0.17	0.71
Race/ethnicity						
White, non-Hispanic	5,225	0.02	0.66	442	-0.21	0.66
Black, non-Hispanic	729	0.17	0.89	201	-0.06	0.85
Hispanic, race specified	696	-0.07	0.75	81	-0.21	0.60
Hispanic, race not specified	740	-0.13	0.71	82	-0.39	0.61
Asian	483	-0.03	0.48	24	-0.13	0.50
Hawaiian, other Pacific						
Islander	104	-0.21	0.52	2	‡	ţ
American Indian/Alaska						
Native	134	-0.13	0.59	43	-0.08	0.58
More than one race,						
non-Hispanic	195	0.15	0.66	14	0.26	0.55
Socioeconomic status						
First quintile (lowest)	1,018	-0.20	0.79	254	-0.19	0.74
Second quintile	1,386	-0.04	0.71	190	-0.13	0.67
Third quintile	1,554	0.04	0.70	130	-0.05	0.66
Fourth quintile	1,627	0.08	0.65	93	-0.18	0.73
Fifth quintile (highest)	1,985	0.16	0.56	69	0.18	0.42
School type						
Public school	6,793	0.00	0.73	798	-0.18	0.71
Private school	1,480	0.10	0.47	78	0.22	0.44

Table 6-11.	Score breakdown, Self-Concept, by eighth-graders, sixth- and seventh-graders, and
	population subgroup: School year 2006–07

† Not applicable.

¹ Reporting standards not met. ¹ Standard deviation.

NOTE: Items were standardized to a mean of 0 and a standard deviation of 1. Scale scores are averages of the respective standardized item scores. Table estimates are based on C7CW0 weight. Subgroup counts do not sum to total because demographic variables are missing for some cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

	Eight	h-graders		Sixth- and	seventh-	graders
Characteristic	Number	Mean	SD^1	Number	Mean	SD
Total sample	8,310	0.01	0.61	886	-0.29	0.68
Sex						
Male	4,069	0.00	0.63	554	-0.29	0.71
Female	4,241	0.02	0.60	332	-0.25	0.64
Race/ethnicity						
White, non-Hispanic	5,225	0.06	0.57	441	-0.16	0.60
Black, non-Hispanic	729	0.00	0.85	200	-0.43	0.83
Hispanic, race specified	695	-0.11	0.68	81	33	0.74
Hispanic, race not specified	739	-0.16	0.66	81	-0.31	0.64
Asian	482	0.00	0.44	24	-0.32	0.50
Hawaiian, other Pacific						
Islander	104	-0.22	0.43	2	‡	†
American Indian/Alaska						
Native	134	-0.24	0.58	42	-0.34	0.62
More than one race, non-						
Hispanic	194	0.12	0.55	14	-0.00	0.73
Socioeconomic status						
First quintile (lowest)	1,017	-0.24	0.71	252	-0.38	0.71
Second quintile	1,384	-0.07	0.63	189	-0.30	0.60
Third quintile	1,554	0.05	0.61	130	-0.23	0.65
Fourth quintile	1,627	0.05	0.57	92	-0.15	0.63
Fifth quintile (highest)	1,985	0.18	0.50	69	0.29	0.46
School type						
Public school	6,791	-0.01	0.65	794	-0.30	0.70
Private school	1,478	0.15	0.39	78	0.08	0.52

Table 6-12.	Score breakdown, Locus of Control, by eighth-graders, sixth- and seventh-graders, and
	population subgroup: School year 2006–07

† Not applicable.

[‡] Reporting standards not met. ¹ Standard deviation.

NOTE: Items were standardized to a mean of 0 and a standard deviation of 1. Scale scores are averages of the respective standardized item scores. Table estimates are based on C7CW0 weight. Subgroup counts do not sum to total because demographic variables are missing for some cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

7. PSYCHOMETRIC CHARACTERISTICS OF THE INDIRECT MEASURES

Chapter 7 describes the selection and development of the eighth-grade indirect measures. The indirect measures were teacher evaluations of children's academic skills in English, mathematics, and science. This chapter provides details on the psychometric characteristics of these instruments. In addition, relationships between the direct and indirect cognitive measures are explored. Weighted means of the indirect measures are presented to provide estimates of child-level assessment scores for eighth grade that are representative of the population cohort. Estimates of the scales of the psychometric characteristics of the indirect measures (e.g., alpha coefficients) are calculated unweighted to examine the characteristics of these data (rather than the population).

7.1 Teacher Measures

In the spring-eighth grade data collection (round 7), teachers of the sampled children were asked to evaluate each child's academic skills by completing questionnaires, based on the Academic Rating Scale (ARS), that evaluated children's skills in three domains. Each teacher received at least one of the three child-level questionnaires (English, mathematics, or science, based on the subject(s) taught by the teacher) for the focal child. Teachers were instructed to rate the child's current skills according to grade-level expectations.

All children were assigned to have an English teacher complete a questionnaire. In fifth grade, half of the children were randomly assigned to have a mathematics teacher complete a questionnaire, and the other half of the children were assigned to have a science teacher complete a questionnaire. This assignment for the mathematics or science teacher questionnaire made in fifth grade was carried forward in eighth grade so that the same children who had a mathematics teacher questionnaire in fifth grade would have a mathematics teacher questionnaire in eighth grade, and those with a science teacher questionnaire in fifth grade teacher taught the sampled child English, mathematics, and science, the teacher was asked to complete an English questionnaire and either a mathematics or science questionnaire, depending upon the domain for which the child was sampled in the fifth grade. In grade eight, 4,661 children had a mathematics teacher complete a questionnaire while 4,672 had a science teacher complete a questionnaire.

The resulting eighth-grade scores use different items and are scaled differently from the measures collected earlier. Eighth-grade scores should not be directly compared with kindergarten through fifth-grade scores with the intent of evaluating gains over time. Data collected in the earlier rounds may, however, be used as covariates in analyzing eighth-grade achievement and behavioral data. Details of the kindergarten, first-grade, third-grade, and fifth-grade teacher measures (and similar behavioral ratings provided by parents) may be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05) (Rock and Pollack 2002), the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062) (Pollack, Rock et al. 2005) and the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062) (Pollack, Rock et al. 2005) and the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062) (Pollack, Rock et al. 2005) and the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062) (Pollack, Rock et al. 2005) and the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Fifth Grade* (NCES 2006–036rev) (Pollack, Atkins-Burnett et al. 2005).

Differential item functioning (DIF) analysis of the Academic Rating Scale (ARS) was considered inappropriate for the rating scale measures. This is because the ratings were produced by the teacher, not by direct observation of tasks performed by the child. Thus, there is a confounding source of difference, namely the teacher's attitudes or potential bias that cannot be separated from the child's performance. Factor analysis of the ARS scales found a very strong first factor, which suggests that a "halo" effect is operating. This suggests that DIF analysis using the total ARS score as the criterion would probably find no evidence of DIF simply because a teacher who rated a child high on one item would tend to rate the same child high on all items. It was probably not the *items* that were functioning differently, but it may have been *teachers* differentially rating children. This would not be a psychometric characteristic of the scale itself.

It is possible that the interaction between teachers' attitudes and demographic characteristics, and the demographic characteristics, cognitive ability, and behavior of children may influence the academic ratings assigned to children. Secondary analysis of these relationships may reveal differences in the standards used in the academic (ARS) ratings.

7.1.1 Indirect Cognitive Assessment Using the Academic Rating Scale (ARS)

The ARS evaluated skills in three domains: English (specifically Oral Expression and Writing skills), mathematics, and science. For each of the scales, the child's primary teacher in the area completed the ratings.

English, mathematics, and science teachers were asked to rate each sampled child on his or her skills in areas relevant to the subject taught. English teachers were asked about children's skills in written and oral expression. Mathematics teachers were asked about children's skills in mathematics, such as problem solving and demonstrating mathematical reasoning. Science teachers were asked about children's skills in science, such as designing an experiment to solve a scientific question and writing a report and preparing a presentation of scientific data. Teachers rated each child's skills, knowledge, and behaviors as "Outstanding (5)," "Very Good (4)," "Good (3)," "Fair (2)," or "Poor (1)." If a skill, knowledge, or behavior had not been introduced into the classroom yet, or if the teacher otherwise did not have the opportunity to observe the skill, the teacher was able to code that item as "Not Applicable/Not Observed." In eighth grade, many schools are departmentalized so different teachers may be rating the child on science and mathematical thinking.

The teacher (ARS) ratings overlap and augment the information gathered through the direct cognitive assessment battery. Although the direct and indirect instruments measure children's skills and behaviors within the same broad curricular domains with some intended overlap, several of the constructs they were designed to measure differ in significant ways. Most important, the teacher rating scales include items designed to measure both the process and products of children's learning in school, whereas the direct cognitive battery is more limited. Because of time and space limitations, the direct cognitive battery is less able to measure the process of children's thinking, including how they express their ideas, solve mathematical problems, or investigate scientific phenomena. The language and literacy teacher ratings collect information on children's oral expression and written composition, areas not assessed on the direct measure.

These criterion-referenced indirect measures are targeted to the specific grade level of the child and draw upon the daily observations made by teachers of the children in their class.

Item response theory (IRT) analysis, using a generalized partial credit model (Muraki 1992), was used to create measures of the reported performance of children on a hierarchy of skills, knowledge, and behavior. The generalized partial credit model, as implemented in the SSI Parscale computer program, uses the pattern of ratings on items to obtain an estimate of the difficulty of each item and to place each child on an interval scale set with a minimum score of one and a maximum score of five. The analysis showed that the reliability of the estimates of the child's ability was very high for all domains (see table 7-1).

Scale	Reliability
English Writing skill ratings	.96
English Oral Expression skill ratings	.93
Mathematics skill ratings	.95
Science skill ratings	.95

Table 7-1.Teacher rating scale reliability statistics for the IRT-based score, spring-
eighth grade: School year 2006–07

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

As mentioned earlier, the teacher rating scores are scaled to have a low value of one and a high value of five to correspond to the 5-point rating scale that teachers used in rating children on these items. Item difficulties and child scores are placed on a common scale. Children had a high probability of receiving a high rating on items whose difficulty was below their scale score, and a lower probability of receiving a high rating on items above their scale score. Therefore, the scores received on the subscales should be interpreted as the child's average item score. Bayesian estimation techniques allow children who received maximum ratings on all the items or minimum ratings on all the items to receive a rating score.

The weighted means and standard deviations for the eighth-grade (T7) teacher rating scores are shown in table 7-2. Score breakdowns for population subgroups are presented in tables 7-13 through 7-16 at the end of this chapter. The items and the metric for the eighth-grade teacher ratings are different from the Academic Rating Scale ratings in earlier rounds of data collection, so the scores are not directly comparable to those for kindergarten, first, third, or fifth grades. With different items used across the grades and separate calibrations performed, the scoring differs from one grade to another.

Table 7-2.Teacher rating scale means and standard deviations, spring-eighth grade: School year2006–07

Scale	Weighted mean	Standard deviation
English Writing skill ratings	2.40	1.30
English Oral Expression skill ratings	2.73	1.20
Mathematics skill ratings	2.48	1.17
Science skill ratings	2.38	1.28

NOTE: Table estimates are based on C7CW0 weight. The range of possible values is 1-5.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

7.1.1.1 Floor and Ceiling

As noted in the section on the development of the ARS, the criteria for some of the items were set very high to avoid serious ceiling problems, and some items were included at a level designed to avoid most floor problems. Because teachers could not be expected to respond to items far outside the range of grade-level performance (they would have little opportunity to observe this as well), it was unavoidable in this type of measure that some children would have perfect scores. Table 7-3 presents the percentage of children at the ceiling and floor of the measures.

Table 7-3.	Percent of sample with perfect and minimum teacher rating scores,
	spring-eighth grade: School year 2006–07

Description	Percent
Perfect scores	
English Writing skill ratings	9.0
English Oral Expression skill ratings	11.1
Mathematics skill ratings	7.0
Science skill ratings	7.6
Minimum scores	
English Writing skill ratings	7.6
English Oral Expression skill ratings	4.6
Mathematics skill ratings	6.2
Science skill ratings	7.8

NOTE: Statistics are unweighted.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

Tables 7-4 to 7-7 provide estimates of difficulty for each of the items. Higher values imply that teachers rated fewer children as proficient on those items. Children would have a greater than 50 percent probability of receiving ratings of "5" on items below their ability level.

Table 7-4.	English Oral Expression item difficulties (arranged in order of difficulty), spring-eighth grade	e:
	School year 2006–07	

Item difficulty	Item number and abbreviated content
2.19	Q12a. Uses Spoken English Grammar
2.61	Q12c. Expresses Creative Thinking
2.72	Q12b. Expresses Analytical or Critical Thinking

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

Table 7-5.English Writing Skills item difficulties (arranged in order of difficulty), spring-eighth grade:
School year 2006–07

Item difficulty	Item number and abbreviated content
2.40	Q11a. Organizes Ideas Logically and Coherently
2.46	Q11c. Gathers Information for Research Purposes
2.46	Q11b. Employs English Grammar and Usage
2.53	Q11d. Writes Various Types of Composition
2.85	Q11e. Uses Style and Rhetoric

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

Table 7-6.Mathematics Skills item difficulties (arranged in order of difficulty), spring-eighth grade:
School year 2006–07

Item difficulty	Item number and abbreviated content
1.48	Q11f. Uses Calculator to Solve Problems
2.23	Q11g. Uses Computer to Complete Mathematics Assignments
2.68	Q11a. Applies Mathematical Concepts to Real World
2.68	Q11c. Talks about Reasoning in Solving a Problem
2.74	Q11e. Uses Representations to Model Mathematical Ideas
2.82	Q11d. Explains Reasoning in Solving a Problem in Writing
2.85	Q11b. Conducts Proofs or Demonstrates Mathematical Reasoning

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99, spring 2007.

Table 7-7. Science Skills item difficulties (arranged in order of difficulty), spring-eighth grade: School year 2006–07

Item difficulty	Item number and abbreviated content
2.33	Q11a. Organizes Data in Tables and Charts
2.50	Q11f. Applies Science Concepts to Solve Real World Problems
2.52	Q11c. Talks about Investigations to Solve Problems
2.57	Q11b. Writes Up Results or Presentation for Research Project
2.64	Q11d. Makes Presentation to Class about Science Analysis
2.79	Q11e. Designs Experiment to Solve Scientific Question

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99, spring 2007

The teacher ratings scales were designed to provide information on children's abilities at a given point in time, not necessarily over time. Moreover, these teacher rating scales are placed on a different score metric than the ARS scores in previous rounds. Therefore, change scores cannot be calculated between time points.

The teacher ratings do not represent a systematic national sample of teachers. Each set of teacher ratings is linked to a sampled child, and teachers were asked to rate each of the ECLS-K sample children they had in class.

Tables 7-8 to 7-11 provide standard errors (SE) for ARS scores for eighth grade. The "Score" column is the sum of the raw score ratings. "Measure" is the average score estimated using the Rating Scale model for the subsample of children who had the corresponding raw score sum. The column labeled "SE" is the corresponding standard error of measurement for those scores. These standard errors can be used in analytic models to correct for the heteroskedasticity of scores.

Table 7-8. English Oral Expression standard errors, spring-eighth grade: School year 2006–07

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
3	1.35	.06	8	2.91	.01	13	4.00	.00
4	1.34	.05	9	3.00	.00	14	4.67	.02
5	2.04	.02	10	3.10	.01	15	5.00	.00
6	2.05	.01	11	3.96	.01			
7	2.07	.01	12	4.00	.00			

NOTE: The "Score" column is the sum of the raw score ratings. "Measure" is the IRT-based score. The column labeled "SE" is the corresponding standard error of measurement for those scores.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99, spring 2007.

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
5	1.11	.02	12	2.45	.03	19	4.05	.01
6	1.41	.04	13	3.00	.00	20	4.03	.01
7	1.44	.03	14	3.05	.01	21	4.00	.00
8	2.14	.02	15	3.08	.01	22	4.00	.00
9	2.14	.02	16	3.19	.02	23	4.69	.03
10	2.10	.01	17	3.11	.02	24	5.00	.00
11	2.25	.02	18	4.00	.01	25	5.00	.00

Table 7-9. English Writing Skills standard errors, spring-eighth grade: School year 2006–07

NOTE: E = estimated extreme score. The "Score" column is the sum of the raw score ratings. "Measure" is the IRT-based score. The column labeled "SE" is the corresponding standard error of measurement for those scores.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
7	1.36	.06	17	2.89	.04	27	4.06	.02
8	1.36	.05	18	3.10	.02	28	4.26	.04
9	1.79	.05	19	3.17	.03	29	4.43	.06
10	1.95	.05	20	3.24	.03	30	4.57	.05
11	2.07	.03	21	3.28	.03	31	4.07	.04
12	2.20	.04	22	3.56	.04	32	4.52	.10
13	2.27	.03	23	3.62	.04	33	5.00	.00
14	2.29	.04	24	3.83	.03	34	5.00	.00
15	2.77	.04	25	4.07	.03	35	5.00	.00
16	2.82	.03	26	3.99	.01			

Table 7-10. Mathematics Skills standard errors, spring-eighth grade: School year 2006–07

NOTE: The "Score" column is the sum of the raw score ratings. "Measure" is the IRT-based score. The column labeled "SE" is the corresponding standard error of measurement for those scores.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
7	1.23	.05	16	3.09	.02	25	4.08	.03
8	1.51	.06	17	3.05	.02	26	4.00	.00
9	2.09	.02	18	3.11	.02	27	4.03	.02
10	2.08	.03	19	3.17	.03	28	5.00	.00
11	2.22	.04	20	3.29	.03	29	5.00	.00
12	2.23	.03	21	3.40	.03	30	5.00	.00
13	2.33	.04	22	4.00	.00			
14	2.25	.03	23	4.08	.03			
15	3.02	.03	24	4.01	.01			

Table 7-11. Science Skills standard errors, spring-eighth grade: School year 2006–07

NOTE: The "Score" column is the sum of the raw score ratings. "Measure" is the IRT-based score. The column labeled "SE" is the corresponding standard error of measurement for those scores.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

7.2 Discriminant and Convergent Validity of the Direct and Indirect Measures

As indicated earlier, the patterns of correlations among selected measures provide evidence for their construct validity, that is, whether they measure what they purport to measure. Systematic evidence for construct validity is often described in terms of *convergent* and *discriminant* validity. Convergent validity means that two different measures of the *same* trait or skill ought to have relatively high correlations with each other. Conversely, discriminant validity means that two measures that are designed to measure two *different* traits or skills should show lower correlations with each other than each does with its matching measure. (An exception to this model is high correlations that may be found for different measures that constitute a predictive relationship.) More complete discussions of construct validity may be found in Campbell and Fiske (1959) and Campbell (1960).

Correlations among 11 eighth-grade measures listed below in exhibit 7-1were examined for evidence of convergent and discriminant validity. These measures included four teacher ratings of children's achievement (ARS), two children's self-ratings of achievement (self-description questionnaire [SDQ])), two selected children's self-perception ratings, and direct cognitive scores in the three subject areas assessed. These correlations are shown in table 7-12. Correlation coefficients with the eighth-grade data were calculated unweighted to be comparable with those from the fifth-grade data.

Exhibit 7-1. Eighth-grade indirect and direct cognitive and noncognitive measures, examined for evidence of convergent and discriminate validity: School year 2006–07

- 1. ARS Oral (Teacher ARS score for English Oral Expression)
- 2. ARS Write (Teacher ARS score for English Writing Skills)
- 3. ARS Math (Teacher ARS score for Mathematics)
- 4. ARS Sci (Teacher ARS score for Science)
- 5. SDQ Read (Child's self-rating of competence in reading)
- 6. SDQ Math (Child's self-rating of competence in mathematics)
- 7. Locus (Child's self-rating of locus of control)
- 8. Concept (Child's self-rating of self-concept)
- 9. ReadTheta (Direct cognitive test theta (ability) estimate for Reading)
- 10. MathTheta (Direct cognitive test theta (ability) estimate for Mathematics)
- 11. SciTheta (Direct cognitive test theta (ability) estimate for Science)

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

Indirect ARS Oral, ARS Write, ARS Math, and ARS Sci measures have counterparts in measures Read Theta, Math Theta, and Science Theta, the direct cognitive assessment scores. It is instructive to compare the discriminant validity within each of the two sets of cognitive measures (the

extent to which scores measuring different constructs should be different), as well as the convergent validity across sets (the extent to which scores should be closely related to other measures of the same construct).

					F	Round 7					
	ARS	ARS	ARS	ARS	SDQ	SDQ			Read	Math	Sci
Measures	Oral	Write	Math	Sci	Read	Math	Locus	Concept	Theta	Theta	Theta
ARS Oral	1.00										
ARS Write	0.84	1.00									
ARS Math	0.51	0.58	1.00								
ARS Sci	0.57	0.60	(1)	1.00							
SDQ Read	0.38	0.42	0.22	0.32	1.00						
SDQ Math	0.17	0.20	0.42	0.28	0.11	1.00					
Locus	0.29	0.33	0.30	0.32	0.26	0.22	1.00				
Concept	0.21	0.23	0.24	0.27	0.24	0.27	0.59	1.00			
Read Theta	0.53	0.58	0.50	0.55	0.35	0.14	0.36	0.24	1.00		
Math Theta	0.49	0.54	0.57	0.54	0.19	0.30	0.32	0.23	0.73	1.00	
Sci Theta	0.48	0.52	0.51	0.51	0.23	0.18	0.32	0.22	0.77	0.78	1.00

Table 7-12.Intercorrelations among the indirect cognitive teacher ratings (ARS), selected child self-
ratings (SDQ, Locus, Concept), and direct cognitive test scores, spring-eighth grade: School
year 2006–07

¹Children were rated by teachers on the ARS mathematics or the ARS Science, but not both. This cell is empty. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

Correlations among the direct cognitive measures are generally similar to those observed in fifth grade. In eighth grade, the correlations for the direct cognitive measures are .73 for the relationship between reading with mathematics and .77 between reading and science. In fifth grade, the corresponding correlation coefficients were .75 and .77, respectively. While the direct cognitive mathematics and science measures were read to the children in fifth grade to remove as much of the reading demands as possible from the content areas, the direct assessments were proctored in eighth grade. As a result, children read and responded to the assessment questions on their own. It was expected that, by eighth grade, children would be very familiar with taking proctored tests. However, it is possible that children with stronger literacy skills might have the opportunity to read more widely in content areas, and the increased exposure to mathematics and science content might increase the development of the concepts and vocabulary needed for success in the content areas.

From kindergarten through the third-grade data collection, the corresponding correlations for ARS were consistently high. In kindergarten through third grade, the same teacher responded to all areas of the ARS (ARS language/literacy, the ARS mathematical thinking, and the ARS science measure).

Thus, there was additional method variance in the correlation. In the fifth grade, the teachers who taught reading, mathematics, and science rated the children on the relevant ARS form; thus, the ARS ratings may have been completed by different teachers. The correlations of the ARS language/literacy with the ARS mathematics scale and with the ARS science scale were lower for the fifth-grade data collection period when compared with previous data collections and when compared with the relationships among the direct measures. The data collection procedures for the ARS in eighth grade were the same as those used in the fifth grade round. However, the correlations of the ARS Oral and ARS Write scales with the ARS Mathematics and ARS Science scales are lower in eighth grade than corresponding correlations found in fifth grade.

When one examines the cross-correlations from a convergent validity perspective, patterns are similar to those found in fifth grade. Relationships are stronger within measures than across measures of similar constructs. One would expect that the direct score in each subject area would be more closely related to the indirect measure of the same subject than to measures of the other subjects. This is true for oral expression and writing skills (ARS with direct reading) and mathematical skills (ARS with direct mathematics), although the differences are relatively small. This represents an improvement in convergent validity compared with kindergarten and first-grade results, where correlations of the ARS mathematical thinking score with the direct cognitive reading score were almost exactly the same as those with the direct mathematics score. In third, fifth, and eighth grade, the ARS science scale was slightly more highly correlated with both reading and mathematics direct scores than it was with the direct science measure that should have been a closer match.

Correlations of children's self-ratings on the self-description questionnaire with other measures in eighth grade are stronger than in the fifth grade. Eighth-grade correlations between the self-description questionnaire measures and the ARS measures ranged from .17 (for the association between SDQ Math and ARS Oral) to .42 (for the associations between SDQ Read and ARS Write and between SDQ Math and ARS Math). Corresponding correlations in fifth grade were lower, ranging from .12 to .27. These observed increases in the strength of the relationships between the children's self-ratings and the ARS measures continue a pattern observed between the third- and fifth-grade data collections. The slightly stronger correlation in eighth grade suggests a continuing increase in children's awareness of their academic performance. Nevertheless, it continues to appear that children use different criteria than teachers use when rating academic competence. Teachers were more knowledgeable about national standards and had more specific criteria to use when rating academic competence. Children's self-perceptions reflected not only the feedback that they received from others about their performance, but

might also have been influenced by self-comparison with peers in their environments. Thus, some children's scores may reflect the "big fish, little pond" phenomenon described by Marsh and his colleagues (Marsh et al. 1995)

See the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Fifth Grade* (NCES 2006–036rev) (Pollack, Atkins-Burnett et al. 2005) for the correlation matrix from fifth grade. As noted earlier, score breakdowns for population subgroups for the indirect measures are presented in tables 7-13 through 7-16.

	Eig	hth-graders		Sixth- an	d seventh-gr	aders
Characteristic	Number	Mean	SD^1	Number	Mean	SD
Total sample	8,047	3.27	1.00	858	2.60	0.93
Sex						
Male	3,942	3.13	1.00	540	2.57	0.93
Female	4,105	3.40	0.98	318	2.67	0.93
Race/ethnicity						
White, non-Hispanic	5,137	3.37	0.95	438	2.79	0.93
Black, non-Hispanic	707	2.93	1.29	189	2.32	0.90
Hispanic, race specified	640	3.19	1.08	78	2.56	0.87
Hispanic, race not specified	694	3.07	1.01	74	2.52	0.86
Asian	455	3.53	0.74	24	3.04	0.73
Hawaiian, other Pacific Islander	100	3.15	0.63	2	‡	†
American Indian/Alaska Native	123	3.14	0.82	39	2.44	0.73
More than one race, non-Hispanic	185	3.53	0.91	13	3.33	1.14
Socioeconomic status						
First quintile (lowest)	954	2.74	1.03	242	2.42	0.89
Second quintile	1,345	3.04	0.99	183	2.57	0.87
Third quintile	1,510	3.30	0.95	124	2.77	0.84
Fourth quintile	1,585	3.44	0.94	93	2.73	1.01
Fifth quintile (highest)	1,946	3.72	0.84	70	3.56	0.83
School type						
Public school	6,585	3.24	1.05	775	2.59	0.94
Private school	1,445	3.47	0.68	77	2.85	0.83

Table 7-13.Score breakdown, English oral expression, by eighth-graders, sixth- and seventh-graders,
and population subgroup: School year 2006–07

† Not applicable.

‡ Reporting standards not met.

¹ Standard deviation.

NOTE: Table estimates are based on C7CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

	Eig	hth-graders		Sixth- an	d seventh-gi	raders
Characteristic	Number	Mean	SD^1	Number	Mean	SD
Total sample	8,040	3.03	1.06	857	2.24	0.95
Sex						
Male	3,939	2.81	1.05	538	2.13	0.92
Female	4,101	3.25	1.02	319	2.45	0.98
Race/ethnicity						
White, non-Hispanic	5,136	3.16	1.02	434	2.43	0.99
Black, non-Hispanic	703	2.64	1.32	191	2.03	0.96
Hispanic, race specified	637	2.91	1.09	78	2.19	0.85
Hispanic, race not specified	695	2.80	1.07	74	2.04	0.65
Asian	456	3.45	0.75	24	2.92	0.77
Hawaiian, other Pacific Islander	100	2.80	0.70	2	‡	ť
American Indian/Alaska Native	123	2.84	0.93	38	1.93	0.78
More than one race, non-Hispanic	184	3.30	0.91	14	2.42	1.12
Socioeconomic status						
First quintile (lowest)	952	2.45	1.03	242	2.04	0.87
Second quintile	1,343	2.74	1.01	185	2.17	0.82
Third quintile	1,509	3.06	0.99	123	2.41	0.94
Fourth quintile	1,584	3.20	1.05	92	2.52	1.08
Fifth quintile (highest)	1,946	3.56	0.90	69	3.33	0.97
School type						
Public school	6,577	3.00	1.11	775	2.23	0.94
Private school	1,446	3.29	0.73	76	2.53	0.83

Table 7-14. Score breakdown, English writing skills, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006-07

† Not applicable.

¹ Reporting standards not met. ¹ Standard deviation.

NOTE: Table estimates are based on C7CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECSL-K), spring 2007.

	Eig	hth-graders		Sixth- an	d seventh-gr	aders
Characteristic	Number	Mean	SD^1	Number	Mean	SD
Total sample	4,002	3.05	0.98	433	2.37	0.90
Sex						
Male	1,962	3.00	0.99	269	2.33	0.89
Female	2,040	3.11	0.97	164	2.45	0.90
Race/ethnicity						
White, non-Hispanic	2,580	3.17	0.92	205	2.70	0.88
Black, non-Hispanic	339	2.75	1.16	90	2.04	0.87
Hispanic, race specified	304	2.89	1.09	47	2.38	0.81
Hispanic, race not specified	342	2.75	1.12	40	1.93	0.65
Asian	229	3.45	0.77	11	2.59	0.39
Hawaiian, other Pacific Islander	59	2.85	0.58	1	‡	†
American Indian/Alaska Native	64	2.82	0.85	28	2.13	0.85
More than one race, non-Hispanic	82	3.37	0.93	10	2.77	0.90
Socioeconomic status						
First quintile (lowest)	440	2.63	1.06	130	2.26	0.93
Second quintile	657	2.89	0.95	90	2.29	0.80
Third quintile	767	3.02	0.92	59	2.41	0.52
Fourth quintile	842	3.21	0.91	52	2.62	1.09
Fifth quintile (highest)	941	3.45	0.90	39	3.04	0.76
School type						
Public school	3,269	3.04	1.03	398	2.35	0.89
Private school	724	3.17	0.74	34	2.69	0.95

Table 7-15. Score breakdown, mathematics skills, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006-07

† Not applicable.

¹ Reporting standards not met. ¹ Standard deviation.

NOTE: Table estimates are based on C7CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2007.

	Eig	ghth-graders		Sixth- an	d seventh-gi	raders
Characteristic	Number	Mean	SD^1	Number	Mean	SD
Total sample	3,987	2.98	1.05	421	2.40	0.99
Sex						
Male	1,954	2.87	1.05	269	2.34	0.95
Female	2,033	3.09	1.03	152	2.51	1.05
Race/ethnicity						
White, non-Hispanic	2,530	3.17	0.98	227	2.49	0.95
Black, non-Hispanic	354	2.42	1.18	88	2.27	1.28
Hispanic, race specified	330	2.70	1.12	34	2.63	0.97
Hispanic, race not specified	345	2.73	1.05	36	2.09	0.69
Asian	219	3.57	0.77	14	2.77	0.59
Hawaiian, other Pacific Islander	44	2.77	0.67	2	*	†
American Indian/Alaska Native	62	3.14	0.85	16	2.44	0.56
More than one race, non-Hispanic	100	3.25	0.84	4	2.00	0.62
Socioeconomic status						
First quintile (lowest)	476	2.31	1.04	112	1.95	0.79
Second quintile	687	2.75	1.00	92	2.60	1.13
Third quintile	748	3.00	0.97	59	2.57	0.81
Fourth quintile	735	3.19	0.92	52	2.57	0.93
Fifth quintile (highest)	988	3.55	0.89	39	3.35	0.95
School type						
Public school	3.251	2.94	1.10	372	2.40	1.01
Private school	728	3.30	0.73	44	2.54	0.85

Table 7-16. Score breakdown, science skills, by eighth-graders, sixth- and seventh-graders, and population subgroup: School year 2006-07

† Not applicable.

‡ Reporting standards not met.

¹ Standard deviation.

NOTE: Table estimates are based on C7CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2007.

This page is intentionally left blank.

REFERENCES

- American Association for the Advancement of Science. (1995). *Benchmarks for Science Literacy*. [on-line]. Available: www.project2061.org.
- Atkins-Burnett, S., and Meisels, S. J. (2001). Measures of Socio-emotional Development in Middle Childhood. (NCES 2001–03). Working Paper. National Center for Education Statistics, U.S. Department of Education. Washington, DC.
- Campbell, D.T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, 15: 546–53.
- Campbell, D.T., and Fiske, D.W. (1959). Convergent and discriminant validation by the multitraitmultimethod matrix. *Psychological Bulletin*, 56: 81–105.
- Cole, N.S., and Moss, P.A. (1989). Bias in Test Use. In R.L. Linn (Ed.), *Educational Measurement*, (3rd Ed., pp. 201–219). New York: American Council on Education/Macmillan.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, *39*: 1–38.
- Dorans, N.J. and Kulick, E. (2006). Differential Item Functioning on the Mini-Mental State Examination: An application of the Mantel-Haenzel and Standardization procedures. *Medical Care, 44*(11) *Suppl3*: S107–S114.
- Ellis, L. A., Marsh, H. W., and Richards, G. E. (2002). A Brief Version of the Self Description Questionnaire II. In R. G. Craven, H. W. Marsh and K. B. Simpson (Eds.), *Proceedings of the 2nd International Biennial Conference Self-Concept Research: Driving International Research Agenda.* Sydney: University of Western Sydney, SELF Research Centre.
- Ferguson, R. F. (1998). Can Schools Narrow the Black-White Test Score Gap? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.
- Gresham, F., and Elliot, S. (1990). Social Skills Rating System. Circle Pines, MN: American Guidance Services, Inc.
- Grissmer, D., Flanagan, A., and Williamson, S. (1998). Why Did the Black-White Score Gap Narrow in the 1970s and 1980s? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.
- Harcourt Brace. (1995). Science Anytime. Orlando, FL: Author.
- Holland, P.W., and Thayer, D.T. (1986). *Differential item function and the Mantel-Haenszel procedure*. (ETS Research Report No. 86-31). Princeton, NJ: ETS.
- Holt. (1986). Science. New York: Author.
- Ingels, S.J., Abraham, S.Y, Karr, R., Spencer, B.D., and Frankel, M.R. (1990). National Education Longitudinal Study of 1988 (NELS:88) Base-Year: Student Component Data File User's Manual.

(NCES 90–464). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

- Kirsch, I.S., Jungblut, A., Jenkins, L., and Kolstad, A. (1993). Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey. (NCES 1993–275). National Center for Education Statistics, U.S. Department of Education. Washington, DC.
- Jencks, C. (1998). Racial Bias in Testing. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.
- Jencks, C., and Phillips, M. (1998). The Black-White Test Score Gap: An Introduction. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.
- Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Publishers.
- Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22: 719–748.
- Marsh, H.W. (1992a). Self Description Questionnaire: A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self-concept. A test manual and a research monograph. Sydney: University of Western Sydney, SELF Research Centre. (Original work published in 1988)
- Marsh, H.W. (1992b). Self Description Questionnaire (SDQ) II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept. An interim test manual and a research monograph. Sydney: University of Western Sydney, SELF Research Centre. (Original work published in 1990)
- Marsh, H.W., Chessor, D., Craven, R., and Roche, L. (1995). *The effect of gifted and talented programs* on academic self-concept: The big fish strikes again. American Educational Research Journal, 32(2): 285–319.
- Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika*, 49: 359–381.
- 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, IL: Scientific Software, Inc.
- Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, 16(2): 159–176.

National Academy of Sciences. (1995). National Science Education Standards. Washington, DC: Author.

- National Assessment Governing Board (NAGB). (1994a). *Reading Framework for the 1994 National* Assessment of Educational Progress. Washington, DC: U.S. Government Printing Office.
- National Assessment Governing Board (NAGB). (1994b). *Geography Framework for the 1994 National* Assessment of Educational Progress. Washington, DC: U.S. Government Printing Office.
- National Assessment Governing Board (NAGB). (1996a). *Mathematics Framework for the 1996 National* Assessment of Educational Progress. Washington, DC: U.S. Government Printing Office.
- National Assessment Governing Board (NAGB). (1996b). Science Framework for the 1996 National Assessment of Educational Progress. Washington, DC: U.S. Government Printing Office.
- National Council of Teachers of Mathematics. (1989). Curriculum and Evaluation Standards for School Mathematics. Reston, VA: Author.
- Phillips, M., Crouse, J., and Ralph, J. (1998). Does the Black-White Test Score Gap Widen After Children Enter School? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.
- Phillips, M., Brooks-Gunn, J., Duncan, G.J., Klebanov, P., and Crane, J. (1998). Family Background, Parenting Practices, and the Black-White Test Score Gap. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap.* Washington, DC: Brookings Institution Press.
- Pollack, J.M., Atkins-Burnett, S., Najarian, M., and Rock, D.A. (2005). Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Fifth Grade (NCES 2006–036rev). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- Pollack, J. M., Rock, D.A., Weiss, M., and Atkins-Burnett, S. (2005). Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for the Third Grade (NCES 2005–062). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- Rock, D.A., and Pollack, J. (1987). The Cognitive Test Battery. In S.J. Ingels et al., *Field Test Report: National Education Longitudinal Study of 1988 (Base Year)*. Chicago, IL: NORC, University of Chicago.
- Rock, D.A., and Pollack, J. (2002). Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade (NCES 2002–05). National Center for Education Statistics, U.S. Department of Education. Washington, DC.
- Rock, D. A. and Stenner, A. J. (2005). Assessment Issues in the Testing of Children at School Entry. *The Future of Children*, *15*(1).
- Rock, D.A., et al. (1995). *Psychometric Report for the NELS: 88 Base Year Test Battery*. (NCES 95–382). National Center for Education Statistics, U.S. Department of Education. Washington, DC.
- Rosenberg, Morris. (1965). Society and the Adolescent Self-Image. Princeton, New Jersey: Princeton University Press

Rouse, C., Brooks-Gunn, J. and McLanahan, S. (2005). Introducing the Issue. *The Future of Children*, 15(1).

Scott-Foresman. (1994). Discover the Wonder. Glenview, IL: Author.

Silver Burdett & Ginn. (1991). Science Horizons. Lexington, MA: Author.

- Steele, C. M., and Aronson, J. (1998). Stereotype Threat and the Test Performance of Academically Successful African Americans. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap.* Washington, DC: Brookings Institution Press.
- Tourangeau, K., Lê, T. and Nord, C. (2005). Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Fifth-Grade Methodology Report (NCES 2006–037). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- Tourangeau, K., Lê, T., Nord, C., and Sorongon, A.G. (Forthcoming). Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Eighth-Grade Methodology Report (NCES 2009–003). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- Tourangeau, K., Nord, C., Lê, T., Pollack, J.M., and Atkins-Burnett, S. (2006). Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Combined User's Manual for the ECLS-K Fifth-Grade Data Files and Electronic Codebooks (NCES 2006–032). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- Tourangeau, K., Nord, C., Lê, T., Sorongon, A.G., and Najarian, M. (Forthcoming). Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K–8 Full Sample Data Files and Electronic Codebooks (NCES 2009– 004). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- Tourangeau, K., Nord, C., Lê, T., Wan, S., Bose, J. and West, J. (2002). Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), User's Manual for the ECLS-K Longitudinal Kindergarten–First Grade Public-Use Data File and Electronic Codebook (NCES 2002–149). National Center for Education Statistics, U.S. Department of Education. Washington, DC.
- U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, *National Assessment of Educational Progress Website Data Tool*, retrieved February 28, 2005 from http://nces.ed.gov/nationsreportcard/naepdata/.
- Wright, B.D., and Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago, IL: MESA Press.
- Yamamoto, K., and Mazzeo, J. (1992). Item Response Theory: Scale linking in NAEP. Journal of Education Statistics, 17: 155–173.

APPENDIX A

SCORE STATISTICS FOR DIRECT COGNITIVE MEASURES BY ROUND OF DATA COLLECTION AND SELECTED CHARACTERISTICS

Table A1. Reading assessment, unweighted sample sizes: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

Total sample	Kouna I	Kound 2	Kound 3	Kound 4	c punox	Round 6	Round 7
	17,625	18,937	5,053	16,336	14,280	11,265	9,225
Sex							
Male	8,984	9,688	2,556	8,349	7,223	5,669	4,627
Female	8,640	9,247	2,497	7,987	7,057	5,596	4,598
Race/ethnicity							
White, Non-Hispanic	10,433	11,073	2,935	9,435	8,094	6,465	5,663
Black, Non-Hispanic	2,854	2,968	782	2,371	1,845	1,275	942
Hispanic, race specified	1,182	1,315	322	1,233	1,261	1,023	782
Hispanic, race not specified	1,195	1,423	377	1,335	1,321	1,081	822
Asian	897	1,089	257	1,042	957	785	513
Hawaiian, Other Pacific Islander	186	202	93	188	171	144	107
American Indian, Alaska Native	354	344	126	298	232	207	178
More than one race, Non-Hispanic	476	473	152	397	379	269	209
Socioeconomic status							
1st quintile (lowest)	2,594	2,917	753	2,363	1,968	1,702	1,270
2nd quintile	3,271	3,503	925	2,796	2,235	1,914	1,568
3rd quintile	3,470	3,686	266	3,003	2,441	1,991	1,688
4th quintile	3,650	3,909	1,019	3,173	2,693	2,310	1,728
5th quintile (highest)	3,880	4,152	1,159	3,642	3,165	2,534	2,071
School type							
Public school	13,737	14,579	3,809	12,998	11,608	9,188	7,629
Private school	3,888	4,358	1,042	3,279	2,624	2,054	1,571

SUUKUE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

Characteristic	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Total sample	18,636	19,649	5,226	16,641	14,374	11,274	9,285
Sex							
Male	9,479	10,041	2,644	8,506	7,289	5,676	4,669
Female	9,156	9,606	2,582	8,135	7,085	5,598	4,616
Race/ethnicity							
White, Non-Hispanic	10,433	11,071	2,935	9,436	8,123	6,470	5,707
Black, Non-Hispanic	2,855	2,962	781	2,371	1,875	1,276	954
Hispanic, race specified	1,588	1,624	389	1,354	1,267	1,024	784
Hispanic, race not specified	1,800	1,834	486	1,518	1,330	1,082	821
Asian	898	1,089	256	1,042	957	785	512
Hawaiian, Other Pacific Islander	187	202	93	188	172	144	107
American Indian, Alaska Native	354	345	126	298	250	208	182
More than one race, Non-Hispanic	473	472	151	397	380	269	209
Socioeconomic status							
1 st quintile (lowest)	3,269	3,426	895	2,572	2,004	1,709	1,287
2nd quintile	3,429	3,607	942	2,839	2,255	1,917	1,586
3rd quintile	3,546	3,721	1,001	3,017	2,456	1,993	1,695
4th quintile	3,676	3,921	1,023	3,178	2,694	2,308	1,734
5th quintile (highest)	3,893	4,161	1,158	3,644	3,168	2,534	2,073
School type							
Public school	14,702	15,260	3,971	13,292	11,694	9,198	7,688
Private school	3,934	4,389	1,043	3,286	2,632	2,054	1,572

unuveighted cample cizes: School years 1008-00 1000-2000 2001-02 2003-04 and 2006-07 ent Mathematics Table A7

Characteristic	Round 5	Round 6	Round 7
Total sample	14,351	11,270	9,304
Sex			
Male	7,274	5,674	4,682
Female	7,077	5,596	4,622
Race/ethnicity			
White, Non-Hispanic	8,113	6,469	5,720
Black, Non-Hispanic	1,870	1,272	954
Hispanic, race specified	1,264	1,024	785
Hispanic, race not specified	1,328	1,082	824
Asian	956	785	513
Hawaiian, Other Pacific Islander	172	144	107
American Indian, Alaska Native	249	209	182
More than one race, Non-Hispanic	379	269	210
Socioeconomic status			
1st quintile (lowest)	2,001	1,707	1,294
2nd quintile	2,251	1,917	1,590
3rd quintile	2,454	1,994	1,698
4th quintile	2,693	2,308	1,734
5th quintile (highest)	3,165	2,534	2,076
School type			
Public school	11,668	9,193	7,706
Private school	2.634	2.054	1.573

Table A3. Science assessment, unweighted sample sizes: School years 2001–02, 2003–04, and 2006–07

first grade. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2002, spring 2004, and spring 2007.

Reading routing test number right, eighth grade assessment	(range of possible values: 0 to 10): School year 2006–07
Table A4.	

	TITACAT	
Characteristic	Mean	SD^{i}
Total sample	5.95	2.61
Sex		
Male	5.87	2.58
Female	6.04	2.63
Race/ethnicity		
White, Non-Hispanic	6.62	2.49
Black, Non-Hispanic	4.52	2.33
Hispanic, race specified	5.45	2.53
Hispanic, race not specified	4.84	2.36
Asian	6.62	2.48
Hawaiian, Other Pacific Islander	6.15	2.65
American Indian, Alaska Native	4.70	2.42
More than one race, Non-Hispanic	6.47	2.29
Socioeconomic status		
1st quintile (lowest)	4.14	2.30
2nd quintile	5.44	2.45
3rd quintile	6.01	2.37
4th quintile	6.63	2.40
5th quintile (highest)	7.79	1.99
School type		
Public school	5.82	2.61
Private school	7.09	2.27
¹ Standard deviation. NOTE: Table estimates are based on C7CW0 weight. There is no kinde	rgarten/first-grade, thir	d-grade,

or fifth-grade variable for comparison. Subgroup counts do not sum to total sample because demographic variables are missing for some cases. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2007.

	Round 7	
Characteristic	Mean	SD^1
Total sample	5.81	2.63
Sex		
Male	5.89	2.68
Female	5.72	2.58
Race/ethnicity		
White, Non-Hispanic	6.40	2.55
Black, Non-Hispanic	4.24	2.27
Hispanic, race specified	5.49	2.45
Hispanic, race not specified	5.06	2.47
Asian	7.10	2.60
Hawaiian, Other Pacific Islander	5.99	2.53
American Indian, Alaska Native	4.59	2.48
More than one race, Non-Hispanic	6.41	2.62
Socioeconomic status		
1st quintile (lowest)	4.29	2.40
2nd quintile	5.16	2.45
3rd quintile	5.86	2.46
4th quintile	6.52	2.39
5th quintile (highest)	7.55	2.21
School type		
Public school	5.72	2.63
Private school	6.64	2.48
Standard deviation.		
NUTE: Table estimates are based on C/CW0 weight. There is no	cindergarten/first-grade, thir	d-grade,

NotE: Table estimates are based on C7CW0 weight. There is no kindergarten may grow, and the setimates are based on C7CW0 weight. There is no kindergarten may grow, and the or fifth-grade variable for comparison. Subgroup counts do not sum to total sample because demographic variables are missing for some cases. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2007.

	Round	17
Characteristic	Mean	SD^{1}
Total sample	6.38	2.47
Sex		
Male	6.56	2.48
Female	6.18	2.44
Race/ethnicity		
White, Non-Hispanic	7.11	2.25
Black, Non-Hispanic	4.66	2.29
Hispanic, race specified	5.93	2.31
Hispanic, race not specified	5.45	2.29
Asian	7.17	2.30
Hawaiian, Other Pacific Islander	5.89	2.25
American Indian, Alaska Native	5.12	2.43
More than one race, Non-Hispanic	6.81	2.25
Socioeconomic status		
1st quintile (lowest)	4.69	2.34
2nd quintile	5.76	2.38
3rd quintile	6.58	2.22
4th quintile	7.14	2.15
5th quintile (highest)	7.94	1.92
School type		
Public school	6.31	2.47
Private school	7.06	2.34
¹ Standard deviation. NOTE: Table estimates are based on C7CW0 weight. There is no	indergarten/first -grade, th	ird-grade, or

NOTE: Table estimates are based on C7CW0 weight. There is no kindergarten/first -grade, third-grade, fifth-grade variable for comparison. Subgroup counts do not sum to total sample because demographic variables are missing for some cases. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2007.

	Roun	nd 1	Roun	d 2	Roun	id 3	Rour	nd 4	Rour	ıd 5	Roun	9 pi	Roun	7 bi
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	34.83	9.81	45.68	13.51	52.27	17.65	75.88	23.79	124.68	28.43	147.07	27.50	166.51	29.25
Sex														
Male	34.24	9.95	44.68	13.44	50.90	17.46	73.97	24.12	122.21	28.99	145.27	28.19	163.75	30.19
Female	35.45	9.62	46.74	13.50	53.71	17.74	77.93	23.26	127.30	27.58	148.98	26.62	169.50	27.89
Race/ethnicity														
White, Non-Hispanic	36.15	10.04	47.50	13.93	54.81	18.54	80.17	24.07	132.23	26.97	153.85	25.86	174.33	25.90
Black, Non-Hispanic	32.44	7.79	41.98	11.04	47.86	13.88	67.81	20.74	111.89	25.55	134.34	25.91	148.98	28.88
Hispanic, race specified	32.85	9.36	43.87	12.22	50.10	14.97	71.86	21.97	118.15	27.49	141.95	26.98	161.79	29.89
Hispanic, race not specified	31.26	7.30	41.40	10.82	45.35	13.01	66.74	20.02	109.96	26.61	133.58	25.83	153.43	29.71
Asian	39.38	14.43	51.89	18.11	61.36	25.04	83.60	26.10	128.90	26.38	151.40	26.07	176.62	25.57
Hawaiian, Other Pacific Islander	33.75	9.58	42.98	11.17	45.75	13.18	72.03	21.47	117.32	25.20	142.60	25.31	165.08	27.70
American Indian, Alaska														
Native	29.38	6.36	39.39	9.52	39.82	10.04	60.68	18.65	103.70	26.68	128.84	30.87	154.90	28.09
More than one race, Non-												Į		
Hispanic	34.86	11.28	45.52	14.88	51.94	16.18	77.09	24.30	125.76	27.65	152.99	24.67	172.86	26.60
Socioeconomic status														
1st quintile (lowest)	29.76	5.83	39.04	8.83	43.13	11.01	62.51	18.21	105.05	25.27	126.44	26.52	144.08	29.65
2nd quintile	32.49	7.53	42.89	11.30	47.93	13.82	71.62	21.13	118.86	26.21	140.89	25.08	159.86	27.95
3rd quintile	34.06	7.84	45.06	11.31	52.59	16.10	76.08	21.34	126.04	25.23	149.57	23.27	168.61	25.58
4th quintile	36.48	9.75	48.12	13.13	55.46	17.28	80.84	22.54	133.35	25.54	155.32	23.70	175.55	24.88
5th quintile (highest)	40.68	12.66	53.22	17.09	61.60	21.86	89.76	25.64	144.29	23.61	165.81	20.80	186.24	18.86
School type														
Public school	34.03	9.19	44.65	12.73	51.26	17.06	74.26	23.11	123.15	28.45	145.57	27.48	165.07	29.44
Private school	39.16	11.74	51.31	16.03	59.39	19.19	86.69	24.65	136.68	25.21	158.73	24.90	179.83	23.62
¹ Standard deviation.													•	
NUTE: Table estimates are based (on a common scale to sumort com	on cross-seci narisons	tional weight	s within each	round (CIC	wu, czcwu	, cscwu, c	4CW0, C5C	WU, C6CWU	, с/сw0). Е	stimates for k	cindergarten t	through eight	h grade have	been put
SOURCE: U.S. Department of Edu	Ication, Nation	onal Center f	or Education	Statistics, E	arly Childhoc	od Longitudi	nal Study, Ki	ndergarten C	lass of 1998.	99 (ECLS-K), fall 1998, s	spring 1999, f	fall 1999, spri	ing 2000,
spring 2002, spring 2004, and sprin	ıg 2007.				•)	•)		,)		J

Reading IRT scale score K-8 scale (range of nossible values: 0 to 212): School vears 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07 Table A7

A-7

70-0007														
	Roun	id 1	Roun	d 2	Roun	id 3	Rour	ıd 4	Roui	ıd 5	Roun	1d 6	Roun	ld 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	25.55	8.90	35.60	11.82	42.67	14.26	60.30	18.16	97.01	25.01	120.58	25.94	138.70	23.55
Sex														
Male	25.61	9.47	35.75	12.50	42.87	15.24	61.03	19.32	98.99	25.88	122.77	25.97	139.42	24.15
Female	25.47	8.26	35.45	11.04	42.46	13.15	59.53	16.80	94.90	23.88	118.23	25.70	137.91	22.86
Race/ethnicity														
White, Non-Hispanic	27.79	9.19	38.61	11.96	46.20	14.42	64.95	18.31	103.61	23.74	126.95	24.19	144.26	21.36
Black, Non-Hispanic	22.40	6.60	30.87	9.43	37.48	11.97	51.51	14.62	83.13	22.25	105.11	23.95	123.37	23.14
Hispanic, race specified	22.58	7.49	32.11	10.40	39.65	12.43	56.29	16.88	91.33	24.14	116.13	24.74	136.48	21.89
Hispanic, race not specified	20.69	6.55	29.73	9.54	35.17	11.14	52.63	14.20	87.27	22.77	112.54	25.02	131.93	23.80
Asian	29.42	10.84	39.63	13.69	47.18	17.05	63.19	19.71	103.66	26.45	130.80	25.87	148.70	23.09
Hawaiian, Other Pacific Islander	23.77	7.69	32.13	10.04	36.67	10.05	52.47	13.64	90.19	21.38	118.80	23.74	139.75	22.01
American Indian, Alaska	1010	007	12.00	67 0	10 00	1611	1013	0071	CL C0	2010	102 45	13 JC	12050	
More than one race, Non-	21.04	0.07	10.00	c0.6	10.70	10.11	10.10	4C.41	71.70	C0.17	C4.C01	70.02	6C.UCI	06.07
Hispanic	25.51	9.02	35.14	11.17	41.21	12.59	60.58	17.81	98.42	24.56	122.41	25.96	145.31	21.85
Socioeconomic status														
1st quintile (lowest)	20.17	5.90	28.66	8.66	33.99	11.21	50.28	14.60	81.09	21.62	102.62	25.80	123.96	24.35
2nd quintile	23.32	7.06	33.21	10.09	39.33	12.22	56.29	16.14	91.41	22.69	114.35	23.32	132.18	23.50
3rd quintile	25.44	7.52	35.72	10.38	43.09	11.97	60.50	16.30	97.55	22.24	122.25	22.04	140.33	20.43
4th quintile	27.69	8.49	38.47	11.20	45.70	12.60	64.64	16.86	105.00	23.11	128.52	22.28	146.16	19.36
5th quintile (highest)	31.62	10.53	43.14	13.20	52.22	16.01	71.95	18.81	113.99	22.18	138.01	20.16	154.01	16.36
School type														
Public school	24.80	8.48	34.71	11.42	41.81	14.05	59.29	17.95	96.24	25.09	119.52	26.11	137.83	23.82
Private school	29.85	10.01	40.72	12.72	49.31	14.11	67.35	17.58	103.68	23.32	128.92	22.64	146.63	19.26
¹ Standard deviation.			. I								-		1 1	
on a common scale to support com	un cross-seci iparisons.	uonal weigni	S WIUNIN CACI	ו Lound (כוכ	wu, uzuwu	, ראראט, ר	40 mg, 020	vu, Louwu,	U/UWU). E	sumates for k	amuergarten u	nrougn eigni	n graue nave	oeen put
SOURCE: U.S. Department of Edu	ucation, Natio	onal Center 1	for Education	Statistics, E	arly Childhoc	od Longitudi	nal Study, Ki	ndergarten C	lass of 1998.	99 (ECLS-K), fall 1998, s	pring 1999, f	all 1999, spri	ing 2000,
spring 2002, spring 2004, and sprin	ng 2007.				•	1)))

Characteristic	Koun	15	Roune	16	Roune	17
Totol comula	Mean	SD^1	Mean	SD	Mean	SD
1 Utal Sality	49.28	15.13	62.84	16.17	82.24	17.26
Sex						
Male	50.88	15.43	64.67	16.06	83.44	17.55
Female	47.57	14.62	60.89	16.05	80.92	16.84
Race/ethnicity						
White, Non-Hispanic	54.77	13.84	68.36	14.02	87.68	14.78
Black, Non-Hispanic	39.35	12.16	51.67	14.91	69.20	17.02
Hispanic, race specified	43.37	14.02	58.15	15.71	79.17	16.30
Hispanic, race not specified	40.24	12.80	54.90	15.43	75.26	17.48
Asian	49.57	15.56	64.14	17.00	87.58	15.27
Hawaiian, Other Pacific Islander	44.71	13.08	57.01	14.43	78.80	15.76
American Indian, Alaska Native	41.86	13.19	50.92	16.29	73.69	17.46
More than one race, Non- Hispanic	51.45	13.97	66.10	13.07	85.83	13.65
Socioeconomic status						
1st quintile (lowest)	38.16	12.24	50.74	15.20	69.54	17.83
2nd quintile	46.40	13.37	59.17	14.69	78.29	16.62
3rd quintile	50.34	13.10	64.61	13.02	83.51	14.63
4th quintile	54.64	13.42	68.02	13.66	88.03	14.08
5th quintile (highest)	60.43	13.36	73.63	13.50	93.47	11.81
School type						
Public school	48.57	15.05	62.06	16.20	81.62	17.44
Private school	54.65	14.52	69.02	14.70	88.14	14.10

Estimates for third through eighth grade have been put on a common scale to support comparisons. Science was not tested in kindergarten/first grade. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2002, spring 2004, and spring 2007.
	Rou	nd 1	Roun	d 2	Roun	id 3	Roui	ld 4	Rour	id 5	Roun	id 6	Rour	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00
Sex														
Male	49.20	10.03	49.06	10.18	48.99	10.19	49.05	10.44	49.11	10.37	49.37	10.25	49.07	10.05
Female	50.84	96.6	51.01	9.70	51.06	69.6	51.02	9.40	50.95	9.51	50.67	9.68	51.01	9.85
Race/ethnicity														
White, Non-Hispanic	51.68	9.71	51.64	9.54	51.75	9.53	51.87	9.26	52.60	9.20	52.44	9.61	52.68	9.54
Black, Non-Hispanic	47.24	9.19	46.83	9.79	47.26	9.46	46.46	10.55	45.63	9.53	45.43	9.13	44.07	8.50
Hispanic, race specified	47.49	9.94	48.54	9.94	48.93	9.25	48.44	9.88	47.78	9.87	48.13	9.53	48.24	9.59
Hispanic, race not specified	45.62	9.25	46.29	9.97	45.16	10.06	46.17	9.91	44.83	10.06	45.18	9.08	45.48	8.98
Asian	54.16	11.60	54.17	10.52	54.09	11.53	52.77	10.02	51.57	8.84	51.43	9.50	53.75	9.77
Hawaiian, Other Pacific Islander	48.44	10.68	47.76	9.94	45.53	9.24	48.82	8.79	47.70	8.85	48.42	8.85	49.53	9.72
American Indian, Alaska														
Native	42.94	8.88	44.36	9.67	40.25	9.76	42.89	10.24	42.44	10.64	43.57	11.04	45.77	8.45
More than one race, Non- Hispanic	49.66	10.56	49.59	10.20	50.01	9.82	50.51	10.02	50.43	9.64	52.09	9.10	51.94	9.23
Socioeconomic status														
1st mintile (lowest)	LL 24	7 04	44.18	0.05	43 56	0 34	44.01	10 31	43.07	0 07	27 65	0 47	47.68	8 57
2nd quintile	47.45	8.69	47.80	9.53	47.40	9.33	48.45	9.73	48.08	9.41	47.70	8.76	47.40	8.76
3rd quintile	49.49	8.91	49.95	9.07	50.67	9.04	50.57	8.78	50.59	8.58	50.76	8.34	50.20	8.60
4th quintile	52.19	9.36	52.39	8.92	52.52	8.67	52.43	8.32	53.03	8.53	52.87	8.76	52.98	9.14
5th quintile (highest)	56.29	10.19	55.64	9.46	55.44	9.22	55.32	8.20	56.63	7.84	56.95	8.38	57.51	8.65
School type														
Public school	49.11	9.74	49.21	9.85	49.38	9.90	49.37	10.01	49.47	10.07	49.45	9.93	49.47	9.91
Private school	54.83	10.00	54.33	9.71	54.66	8.70	54.30	8.49	54.17	8.32	54.32	9.53	54.91	9.45
¹ Standard deviation.														
NOTE: Table estimates are based a common scale to sumort comman	on cross-sec risons	tional weight	s within each	round (CIC	W0, C2CW0	, c3CW0, C	4CW0, C5CV	V0, C6CW0,	C7CW0). Es	stimates for k	indergarten tl	hrough eightl	h grade have	been put or
SOURCE: U.S. Department of Ed	lucation, Nat	ional Center	for Education	n Statistics, 1	Early Childhe	od Longitud	linal Study, k	Cindergarten	Class of 199	8-99 (ECLS-	K), fall 1998.	spring 1999.), fall 1999, s	pring 2000.
spring 2002, spring 2004, and sprin	ng 2007.				`)	•)		,) -)

Characteristic	Roun	d 1	Round	d 2	Roun	d 3	Roun	d 4	Rour	id 5	Roun	id 6	Rour	d 7
	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00
Sex														
Male	49.90	10.44	49.98	10.34	49.92	10.57	50.18	10.51	50.76	10.38	50.88	10.15	50.38	10.23
Female	50.10	9.52	50.02	9.62	50.08	9.36	49.80	9.42	49.19	9.51	49.06	9.75	49.59	9.73
Race/ethnicity														
White, Non-Hispanic	52.73	9.46	52.69	9.34	52.59	9.17	52.48	9.35	52.61	9.37	52.45	9.53	52.34	9.53
Black, Non-Hispanic	46.46	8.72	45.87	9.24	46.26	9.80	45.21	9.92	44.50	9.21	44.06	8.77	43.60	8.54
Hispanic, race specified	46.39	9.58	46.91	9.77	47.98	9.64	47.88	9.99	47.78	9.75	48.21	9.34	48.74	8.99
Hispanic, race not specified	43.81	9.22	44.58	9.67	44.39	9.73	46.09	9.09	46.15	9.33	46.86	9.27	46.98	9.32
Asian	54.16	10.16	53.25	9.84	52.75	10.27	51.39	9.97	52.61	10.46	54.21	10.54	55.45	11.26
Hawallan, Other Pacific Islander	48.06	9.30	47.09	9.14	46.11	7.85	46.13	8.45	47.34	8.76	49.26	8.98	50.02	8.87
American Indian, Alaska														
Native	44.22	9.58	45.56	9.36	42.05	10.47	45.20	9.29	44.37	8.93	43.61	9.72	46.49	9.35
More than one race, Non- Hisnanic	50.11	9 42	49 79	931	49.20	030	50.15	9 95	50.52	9 83	50.75	10.13	53.07	10.00
		1			1	2							2	
Socioeconomic status														
1st quintile (lowest)	43.22	8.70	43.62	9.11	43.32	9.74	44.40	9.97	43.61	9.15	43.23	9.50	43.88	8.96
2nd quintile	47.59	9.00	48.09	9.38	47.80	9.40	48.00	9.79	47.84	9.08	47.45	8.67	47.05	9.21
3rd quintile	50.33	8.68	50.47	8.82	50.78	8.58	50.46	8.90	50.26	8.72	50.45	8.40	50.28	8.69
4th quintile	52.83	8.85	52.75	8.79	52.59	8.19	52.56	8.49	53.16	9.09	52.95	8.76	53.01	8.88
5th quintile (highest)	56.61	9.35	56.12	9.09	56.22	8.84	55.81	8.27	56.70	8.69	56.96	8.49	57.15	8.69
School type														
Public school	49.15	9.85	49.25	9.92	49.37	10.07	49.46	10.07	49.69	10.06	49.60	10.04	49.63	10.02
Private school	54.87	9.43	54.30	9.34	54.80	7.96	53.84	8.29	52.70	9.06	53.16	8.97	53.40	9.21
¹ Standard deviation.											-			
NULE: Lable estimates are based of	n cross-secti	ional weight	s within each	round (CIC	wu, uzuwu,	, cscwu, c	40.WU, CSCV	vu, പപ്പെ സ	C/CWU). E	stimates for h	cindergarten t	nrough eight	h grade have	been put
on a common scale to support comp SOURCE U.S. Denartment of Educ	arisons. vation, Natio	vnal Center fi	or Education	Statistics, Ea	rlv Childhoo	d I ongitudii	ol Study Kin	dargartan Cl	of 1000		· C-11 1000	7 0001		0000

spring 2002, spring 2004, and spring 2007.

	Roun	d 5	Roun	d 6	Roune	17
Characteristic	Mean	SD^{1}	Mean	SD	Mean	SD
Total sample	50.00	10.00	50.00	10.00	50.00	10.00
Sex						
Male	51.05	10.05	51.14	9.97	50.82	10.32
Female	48.88	9.82	48.79	9.89	49.10	9.56
Race/ethnicity						
White, Non-Hispanic	53.66	8.66	53.38	8.67	53.15	9.19
Black, Non-Hispanic	43.36	8.92	43.17	9.30	42.69	8.65
Hispanic, race specified	46.07	9.73	47.16	9.71	47.92	8.96
Hispanic, race not specified	43.95	9.25	45.14	9.59	45.87	9.25
Asian	50.19	10.09	50.84	10.52	53.38	9.86
Hawaiian, Other Pacific Islander	46.97	9.36	46.43	8.76	47.54	8.26
American Indian, Alaska Native	45.19	9.14	42.46	10.65	45.07	9.25
More than one race, Non- Hispanic	51.56	8.96	51.99	7.87	51.69	8.27
Socioeconomic status						
1st quintile (lowest)	42.46	9.08	42.52	9.65	42.88	9.02
2nd quintile	48.28	8.96	47.80	8.89	47.47	9.07
3rd quintile	50.91	8.34	51.07	7.84	50.28	8.38
4th quintile	53.59	8.38	53.14	8.38	53.28	8.82
5th quintile (highest)	57.14	8.16	56.67	8.60	57.13	8.45
School type						
Public school	49.53	10.01	49.51	10.03	49.64	10.03
Private school	53.55	9.10	53.83	9.07	53.45	9.01

Science T-scores, standardized within round (range of possible values: Table A12.

Estimates for third through eighth grade have been put on a common scale to support comparisons. Science was not tested in kindergarten/first grade. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

	Roun	d 1	Roune	d 2	Roun	d 3	Roun	ld 4	Rour	d 5	Roun	9 P	Roun	ld 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	-1.32	0.51	-0.76	0.50	-0.52	0.51	0.08	0.47	0.77	0.32	1.02	0.30	1.28	0.39
Sex														
Male	-1.36	0.51	-0.80	0.51	-0.57	0.52	0.04	0.49	0.74	0.33	1.00	0.31	1.24	0.39
Female	-1.28	0.50	-0.70	0.49	-0.47	0.50	0.13	0.44	0.80	0.30	1.04	0.29	1.32	0.38
Race/ethnicity														
White, Non-Hispanic	-1.23	0.49	-0.67	0.48	-0.43	0.49	0.17	0.44	0.85	0.29	1.09	0.29	1.38	0.37
Black, Non-Hispanic	-1.46	0.47	-0.91	0.49	-0.66	0.49	-0.08	0.50	0.63	0.30	0.88	0.28	1.05	0.33
Hispanic, race specified	-1.45	0.51	-0.83	0.50	-0.57	0.48	0.01	0.47	0.70	0.31	0.96	0.29	1.21	0.37
Hispanic, race not specified	-1.54	0.47	-0.94	0.50	-0.77	0.52	-0.10	0.47	0.61	0.32	0.87	0.27	1.10	0.35
Asian	-1.11	0.59	-0.55	0.53	-0.31	0.59	0.21	0.47	0.82	0.28	1.06	0.29	1.43	0.38
Hawaiian, Other Pacific	1 40	0 5 4	100	0 5 0	31.0	010		11 0					30.1	000
ısıander American Indian, Alaska	-1.40	40.0	-0.8/	00.0	c/·n-	0.48	cu.u	0.41	0. /0	0.28	16.0	17.0	1.20	8c.U
Native	-1.68	0.45	-1.04	0.48	-1.02	0.50	-0.25	0.48	0.53	0.34	0.82	0.33	1.12	0.33
More than one race, Non- Hispanic	-1.34	0.54	-0.78	0.51	-0.52	0.51	0.11	0.47	0.78	0.31	1.08	0.28	1.35	0.36
Socioeconomic status														
1st quintile (lowest)	-1.64	0.40	-1.05	0.45	-0.85	0.48	-0.20	0.49	0.55	0.32	0.80	0.28	1.00	0.33
2nd quintile	-1.45	0.44	-0.87	0.48	-0.65	0.48	0.01	0.46	0.71	0.30	0.95	0.26	1.18	0.34
3rd quintile	-1.35	0.45	-0.76	0.45	-0.49	0.47	0.11	0.41	0.79	0.27	1.04	0.25	1.29	0.33
4th quintile	-1.21	0.48	-0.64	0.45	-0.39	0.45	0.20	0.39	0.87	0.27	1.10	0.26	1.40	0.35
5th quintile (highest)	-1.00	0.52	-0.47	0.47	-0.24	0.47	0.33	0.39	0.98	0.25	1.23	0.25	1.57	0.34
School type														
Public school	-1.37	0.50	-0.79	0.49	-0.55	0.51	0.05	0.47	0.75	0.32	1.00	0.30	1.26	0.38
Private school	-1.07	0.51	-0.54	0.49	-0.28	0.45	0.29	0.40	0.90	0.26	1.15	0.29	1.47	0.37
¹ Standard deviation. NOTE: Table estimates are based on a common scale to support cor SOURCE: U.S. Department of Ed	on cross-sect nparisons. lucation, Natio	ional weight onal Center f	s within each or Education	round (C1C Statistics, E	W0, C2CW0, arly Childhoo	, C3CW0, C d Longitudi	4CW0, C5CV nal Study, Kii	V0, C6CW0, ndergarten C	C7CW0). E4 lass of 1998-	stimates for l 99 (ECLS-K	cindergarten ti), fall 1998, s	hrough eight pring 1999, f	h grade have fall 1999, spri	been put ing 2000,
spring 2002, spring 2004, and spr	ing 2007.													

	Roun	d 1	Roun	d 2	Roune	d 3	Roun	ld 4	Roun	d 5	Roun	d 6	Roune	1 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	-1.19	0.48	-0.71	0.46	-0.46	0.48	0.03	0.43	0.70	0.39	1.07	0.42	1.41	0.46
Sex														
Male	-1.20	0.50	-0.71	0.48	-0.46	0.50	0.04	0.46	0.72	0.41	1.11	0.43	1.43	0.47
Female	-1.19	0.45	-0.71	0.45	-0.46	0.45	0.02	0.41	0.66	0.37	1.03	0.41	1.39	0.44
Race/ethnicity														
White, Non-Hispanic	-1.06	0.45	-0.59	0.43	-0.34	0.44	0.14	0.41	0.80	0.37	1.18	0.40	1.52	0.43
Black, Non-Hispanic	-1.36	0.42	-0.90	0.43	-0.64	0.47	-0.18	0.43	0.48	0.36	0.82	0.37	1.12	0.39
Hispanic, race specified	-1.36	0.46	-0.86	0.45	-0.56	0.46	-0.06	0.43	0.61	0.38	1.00	0.39	1.35	0.41
Hispanic, race not specified	-1.49	0.44	-0.96	0.45	-0.73	0.46	-0.14	0.40	0.54	0.37	0.94	0.39	1.27	0.43
Asian	-0.99	0.48	-0.56	0.46	-0.33	0.49	0.09	0.43	0.80	0.41	1.25	0.44	1.66	0.51
Hawaiian, Other Pacific Islander	-1.28	0.44	-0.85	0.42	-0.64	0.37	-0.14	0.37	0.59	0.34	1.04	0.38	1.41	0.40
American Indian, Alaska														
Native Marine Marine	-1.47	0.46	-0.92	0.43	-0.84	0.50	-0.18	0.40	0.47	0.35	0.80	0.41	1.25	0.43
INOTE LITAL ONE FACE, INOI- Hispanic	-1.19	0.45	-0.72	0.43	-0.50	0.45	0.04	0.43	0.72	0.39	1.10	0.43	1.55	0.46
Socioeconomic status														
1st quintile (lowest)	-1.51	0.41	-1.01	0.42	-0.78	0.46	-0.21	0.43	0.44	0.36	0.79	0.40	1.13	0.41
2nd quintile	-1.31	0.43	-0.80	0.43	-0.56	0.45	-0.06	0.43	0.61	0.36	0.97	0.36	1.28	0.42
3rd quintile	-1.18	0.41	-0.69	0.41	-0.42	0.41	0.05	0.39	0.71	0.34	1.09	0.35	1.42	0.40
4th quintile	-1.06	0.42	-0.58	0.41	-0.34	0.39	0.14	0.37	0.82	0.36	1.20	0.37	1.55	0.41
5th quintile (highest)	-0.88	0.45	-0.43	0.42	-0.16	0.42	0.28	0.36	0.96	0.34	1.36	0.36	1.74	0.40
School type														
Public school	-1.23	0.47	-0.75	0.46	-0.49	0.48	0.01	0.44	0.68	0.40	1.06	0.42	1.40	0.46
Private school	-0.96	0.45	-0.51	0.43	-0.23	0.38	0.20	0.36	0.80	0.36	1.21	0.38	1.57	0.42
¹ Standard deviation.				i i										
NOTE: Table estimates are based on a common scale to summer com	on cross-sect	ional weight	s within each	Lound (CIC	W0, CZCWU,	വാവം വ	4CWU, CSCV	VU, C6CWU,	C7CWU). ES	timates for k	indergarten ti	hrough eighti	h grade have I	been put
SOURCE: U.S. Denartment of Edu	reation. Natic	nal Center f	or Education	Statistics. Ea	rdv Childhoo	d I ,ongitudii	al Study. Kir	ndergarten Cl	ass of 1998-	99 (ECLS-K)	. fall 1998. si	nring 1999. f	all 1999. sprii	nø 2000.
spring 2002, spring 2004, and sprin	ng 2007.					D		0						p 1

Ţ	Roune	d 5	Roune	d 6	Roune	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD
Total sample	-0.66	0.67	-0.08	0.68	0.86	0.88
Sex						
Male	-0.59	0.68	0.00	0.68	0.93	0.91
Female	-0.74	0.66	-0.16	0.68	0.78	0.84
Race/ethnicity						
White, Non-Hispanic	-0.41	0.58	0.15	0.59	1.14	0.81
Black, Non-Hispanic	-1.11	0.60	-0.55	0.64	0.22	0.76
Hispanic, race specified	-0.93	0.66	-0.27	0.66	0.68	0.79
Hispanic, race not specified	-1.07	0.62	-0.41	0.66	0.50	0.81
Asian	-0.65	0.68	-0.02	0.72	1.16	0.87
Hawaiian, Other Pacific Islander	-0.87	0.63	-0.32	0.60	0.64	0.73
American Indian, Alaska Native More than one race Non-	-0.99	0.62	-0.60	0.73	0.43	0.81
Hispanic	-0.56	09.0	0.06	0.54	1.01	0.73
Socioeconomic status						
1st quintile (lowest)	-1.17	0.61	-0.59	0.66	0.23	0.79
2nd quintile	-0.78	0.60	-0.23	0.61	0.64	0.80
3rd quintile	-0.60	0.56	-0.01	0.54	0.88	0.74
4th quintile	-0.42	0.56	0.13	0.57	1.15	0.78
5th quintile (highest)	-0.18	0.55	0.38	0.59	1.49	0.74
School type						
Public school	-0.69	0.67	-0.11	0.69	0.83	0.88
Private school	-0.42	0.61	0.18	0.62	1.16	0.79

Table A15.Science IRT theta score, 5-8 scale (range of possible values: -5 to 5):School years 2001–02, 2003–04, and 2006–07

NOTE: Table estimates are based on cross-sectional weights within each round (C5CW0, C6CW0, C7CW0). Estimates for third through eighth grade have been put on a common scale to support comparisons. Science was not tested in kindergarten/first grade.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2007.

	Roun	d 1	Round	12	Round	d 3	Roun	id 4	Rour	id 5	Roun	d 6	Roun	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.65	0.34	0.92	0.19	0.96	0.14	66.0	0.05	1.00	0.00	1.00	0.00	1.00	0.00
Sex														
Male	0.62	0.35	0.90	0.20	0.95	0.15	0.99	0.06	1.00	0.00	1.00	0.00	1.00	0.00
Female	0.68	0.33	0.93	0.17	0.97	0.12	1.00	0.04	1.00	0.00	1.00	0.00	1.00	0.00
Race/ethnicity														
White, Non-Hispanic	0.71	0.32	0.94	0.15	0.97	0.11	1.00	0.04	1.00	0.00	1.00	0.00	1.00	0.00
Black, Non-Hispanic	0.57	0.35	0.87	0.22	0.94	0.16	0.99	0.07	1.00	0.00	1.00	0.00	1.00	0.00
Hispanic, race specified	0.56	0.36	0.89	0.21	0.96	0.11	0.99	0.06	1.00	0.00	1.00	0.00	1.00	0.00
Hispanic, race not specified	0.50	0.36	0.85	0.25	0.90	0.21	0.99	0.06	1.00	0.00	1.00	0.00	1.00	0.00
Asian	0.76	0.30	0.96	0.12	0.98	0.09	0.99	0.05	1.00	0.00	1.00	0.00	1.00	0.00
Hawaiian, Other Pacific Islander	0.60	0.36	0.88	0.21	0.94	0.13	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
American Indian, Alaska					-									
Native	0.39	0.35	0.82	0.26	0.82	0.28	0.99	0.06	1.00	0.00	1.00	0.00	1.00	0.00
More than one race, Non- Hispanic	0.63	0.35	0.92	0.18	0.95	0.16	0.99	0.06	1.00	0.00	1.00	0.00	1.00	0.00
6	5) • •) • •			0 0 0)))		0 0 0		•
Socioeconomic status														
1st quintile (lowest)	0.43	0.34	0.83	0.26	0.89	0.22	0.98	0.09	1.00	0.00	1.00	0.00	1.00	0.00
2nd quintile	0.58	0.34	0.89	0.21	0.94	0.16	0.99	0.05	1.00	0.00	1.00	0.00	1.00	0.00
3rd quintile	0.65	0.33	0.93	0.17	0.97	0.12	1.00	0.04	1.00	0.00	1.00	0.00	1.00	0.00
4th quintile	0.74	0.30	0.96	0.12	0.98	0.07	1.00	0.03	1.00	0.00	1.00	0.00	1.00	0.00
5th quintile (highest)	0.83	0.26	0.97	0.10	0.99	0.04	1.00	0.01	1.00	0.00	1.00	0.00	1.00	0.00
School type														
Public school	0.62	0.35	0.91	0.20	0.95	0.14	0.99	0.06	1.00	0.00	1.00	0.00	1.00	0.00
Private school	0.80	0.27	0.96	0.12	0.99	0.06	1.00	0.03	1.00	0.00	1.00	0.00	1.00	0.00
¹ Standard deviation.														
NOTE: Table estimates are based (on cross-sections	ional weight	s within each	round (C1C)	W0, C2CW0,	, C3CW0, C	34CW0, C5CV	V0, C6CW0,	C7CW0). E.	stimates for k	indergarten tl	hrough eighti	h grade have l	been put
SOURCE II S Denartment of Edu	reation Natio	unal Center f	or Education :	Statistics Ea	rlv Childhoo	d Lonoitudi	nal Study Kir	nderoarten Cl	lass of 1998-	99 (FCI S-K)) fall 1998 st	nring 1999 f	all 1999 snri	no 2000
spring 2002, spring 2004, and sprir	ng 2007.	· · · · · · · · · · · · · · · · · · ·	01 Fuuvuu	formoning		munduron n		~	> / Y YA 0001		1, 101 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	- (16 ± < < < ,

Probability of proficiency, reading level 1: letter recognition (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-07, 2003–04, and 2006–07 Table A16.

•	Round	11	Round	12	Round	13	Roun	id 4	Rour	ld 5	Roun	id 6	Roun	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.29	0.33	0.68	0.33	0.81	0.28	96.0	0.14	1.00	0.00	1.00	0.00	1.00	0.00
Sex														
Male	0.27	0.32	0.64	0.35	0.78	0.30	0.95	0.15	1.00	0.00	1.00	0.00	1.00	0.00
Female	0.32	0.34	0.72	0.31	0.84	0.25	0.97	0.12	1.00	0.00	1.00	0.00	1.00	0.00
Race/ethnicity														
White, Non-Hispanic	0.34	0.34	0.74	0.30	0.86	0.24	0.97	0.11	1.00	0.00	1.00	0.00	1.00	0.00
Black, Non-Hispanic	0.21	0.28	0.57	0.35	0.74	0.30	0.93	0.18	1.00	0.00	1.00	0.00	1.00	0.00
Hispanic, race specified	0.23	0.31	0.63	0.35	0.78	0.28	0.96	0.13	1.00	0.00	1.00	0.00	1.00	0.00
Hispanic, race not specified	0.18	0.27	0.56	0.36	0.67	0.34	0.94	0.16	1.00	0.00	1.00	0.00	1.00	0.00
Asian	0.41	0.37	0.78	0.28	0.86	0.23	0.97	0.12	1.00	0.00	1.00	0.00	1.00	0.00
Hawanan, Other Pacific Islander	0.26	0.33	0.59	0.36	0.67	0.32	0.98	0.06	1.00	0.00	1.00	0.00	1.00	0.00
Allicitcali likulati, Alaska Native	0.12	0.23	0.49	0.36	0.51	0.35	0.90	0.19	1.00	0.00	1.00	0.00	1.00	0.00
More than one race, Non- Hispanic	0.28	0.33	0.66	0.33	0.82	0.27	0.96	0.15	1.00	0.00	1.00	0.00	1.00	0.00
Socioeconomic status														
1st quintile (lowest)	0.11	0.20	0.48	0.35	0.63	0.34	0.91	0.20	1.00	0.00	1.00	0.00	1.00	0.00
2nd quintile	0.20	0.27	0.61	0.34	0.75	0.30	0.95	0.15	1.00	0.00	1.00	0.00	1.00	0.00
3rd quintile	0.27	0.30	0.69	0.32	0.84	0.24	0.97	0.10	1.00	0.00	1.00	0.00	1.00	0.00
4th quintile	0.35	0.34	0.76	0.28	0.88	0.20	0.98	0.08	1.00	0.00	1.00	0.00	1.00	0.00
5th quintile (highest)	0.50	0.36	0.84	0.24	0.92	0.16	0.99	0.05	1.00	0.00	1.00	0.00	1.00	0.00
School type														
Public school	0.26	0.32	0.66	0.34	0.79	0.28	0.96	0.14	1.00	0.00	1.00	0.00	1.00	0.00
Private school	0.45	0.36	0.80	0.27	0.92	0.16	0.99	0.07	1.00	0.00	1.00	0.00	1.00	0.00
¹ Standard deviation.		•	-	Č.										
NULE: Lable estimates are based (OII CLOSS-SECU	ional weight.	s within each	רור (רור	wu, uzuwu,	n non non non non non non non non non n	40 M U, U)U	vu, cocwu,	U/UWU). E	sumates for k	amuergarien u	nrougn eigni	n graue nave	oeen put
on a common scale to support com	parisons.	(

Table A17. Probability of proficiency, reading level 2: beginning sounds (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000,

spring 2002, spring 2004, and spring 2007.

	Roun	d 1	Round	12	Roune	d 3	Roun	d 4	Rour	d 5	Roune	d 6	Round	17
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.16	0.26	0.49	0.34	0.65	0.33	0.91	0.20	1.00	0.01	1.00	00.00	1.00	0.00
Sex														
Male	0.15	0.25	0.46	0.35	0.61	0.34	0.90	0.21	1.00	0.01	1.00	0.00	1.00	0.00
Female	0.18	0.26	0.52	0.34	0.68	0.31	0.93	0.17	1.00	0.01	1.00	0.00	1.00	0.00
Race/ethnicity														
White, Non-Hispanic	0.19	0.27	0.55	0.33	0.71	0.30	0.94	0.16	1.00	0.01	1.00	0.00	1.00	0.00
Black, Non-Hispanic	0.10	0.20	0.38	0.34	0.55	0.34	0.85	0.25	1.00	0.01	1.00	0.00	1.00	0.00
Hispanic, race specified	0.12	0.23	0.44	0.34	0.61	0.34	0.90	0.20	1.00	0.01	1.00	0.00	1.00	0.00
Hispanic, race not specified	0.09	0.18	0.37	0.33	0.49	0.35	0.86	0.23	1.00	0.01	1.00	0.00	1.00	0.00
Asian	0.26	0.32	0.61	0.33	0.71	0.32	0.94	0.17	1.00	0.01	1.00	0.00	1.00	0.00
Hawaiian, Other Pacific Islander	0.15	0.25	0.41	0.35	0.47	0.34	0.92	0.15	1.00	0.02	1.00	0.00	1.00	0.00
American Indian, Alaska														
Maria than and man Man	0.06	0.15	0.31	0.32	0.32	0.31	0.79	0.28	0.99	0.02	1.00	0.00	1.00	0.00
Hispanic Hispanic	0.16	0.26	0.46	0.34	0.65	0.31	0.92	0.20	1.00	0.01	1.00	0.00	1.00	0.00
Socioeconomic status														
1st quintile (lowest)	0.05	0.12	0.29	0.30	0.43	0.33	0.82	0.27	0.99	0.02	1.00	0.00	1.00	0.00
2nd quintile	0.10	0.19	0.41	0.33	0.56	0.33	0.90	0.21	1.00	0.01	1.00	0.00	1.00	0.00
3rd quintile	0.14	0.22	0.49	0.33	0.68	0.30	0.94	0.15	1.00	0.01	1.00	0.00	1.00	0.00
4th quintile	0.20	0.27	0.57	0.32	0.74	0.27	0.96	0.13	1.00	0.01	1.00	0.00	1.00	0.00
5th quintile (highest)	0.32	0.33	0.68	0.30	0.81	0.24	0.97	0.09	1.00	0.00	1.00	0.00	1.00	0.00
School type														
Public school	0.14	0.24	0.46	0.34	0.63	0.33	0.91	0.20	1.00	0.01	1.00	0.00	1.00	0.00
Private school	0.27	0.31	0.64	0.32	0.80	0.24	0.96	0.11	1.00	0.00	1.00	0.00	1.00	0.00
¹ Standard deviation.		- - -								۔ ب	· ·		-	
NULE: Table estimates are based (on a common scale to sumort com	on cross-sect	ional weight.	s within each	round (CIC	wu, czcwu,	cscwu, c	4CW0, C5CV	ഗറ, ലംലംഗ,	C/CWU). E	stimates for k	indergarten th	irough eightf	n grade have t	een put
SOURCE: U.S. Department of Edu	ication, Natic	unal Center f	or Education	Statistics, Ea	rlv Childhoo	d Longitudi	nal Study, Kir	ndergarten C	lass of 1998-	99 (ECLS-K)	. fall 1998, sp	ring 1999, fa	all 1999, sprir	ie 2000,
spring 2002, spring 2004, and sprir	ıg 2007.				•)	•)		~	•)	•	D

Probability of proficiency, reading level 3: ending sounds (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02. 2003–04. and 2006–07 Table A18.

2003–04, an	nd 2006–	07												
	Roun	ld 1	Roune	12	Roune	d 3	Roun	d 4	Roun	d 5	Roune	d 6	Roune	17
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.03	0.13	0.15	0.25	0.26	0.32	0.73	0.33	0.98	0.07	1.00	0.01	1.00	0.00
Sex														
Male	0.03	0.13	0.13	0.24	0.24	0.31	0.69	0.35	0.98	0.08	1.00	0.01	1.00	00.00
Female	0.03	0.12	0.16	0.26	0.29	0.33	0.77	0.31	0.99	0.06	1.00	0.01	1.00	0.00
Race/ethnicity														
White, Non-Hispanic	0.04	0.14	0.17	0.27	0.31	0.33	0.79	0.29	0.99	0.05	1.00	0.01	1.00	0.00
Black, Non-Hispanic	0.02	0.09	0.10	0.21	0.19	0.28	0.62	0.37	0.97	0.09	1.00	0.02	1.00	0.00
Hispanic, race specified	0.02	0.10	0.12	0.23	0.23	0.30	0.68	0.35	0.98	0.09	1.00	0.01	1.00	0.00
Hispanic, race not specified	0.01	0.07	0.09	0.19	0.15	0.26	0.60	0.37	0.96	0.10	1.00	0.02	1.00	0.00
Asian	0.08	0.23	0.26	0.34	0.41	0.42	0.80	0.31	0.99	0.04	1.00	0.01	1.00	0.00
Hawaiian, Other Pacific														
Islander	0.03	0.13	0.13	0.23	0.15	0.27	0.68	0.34	0.98	0.09	1.00	0.00	1.00	0.00
American Indian, Alaska														
Native	0.01	0.05	0.06	0.16	0.07	0.17	0.48	0.38	0.94	0.14	0.99	0.02	1.00	0.00
More than one race, Non-														
Hispanic	0.04	0.16	0.14	0.26	0.27	0.33	0.76	0.31	0.99	0.06	1.00	0.00	1.00	0.00
Socioeconomic status														
1st quintile (lowest)	0.00	0.05	0.05	0.14	0.11	0.21	0.53	0.37	0.95	0.13	0.99	0.02	1.00	0.00
2nd quintile	0.01	0.08	0.10	0.21	0.18	0.27	0.69	0.34	0.98	0.08	1.00	0.01	1.00	0.00
3rd quintile	0.02	0.09	0.13	0.23	0.27	0.32	0.76	0.31	0.99	0.05	1.00	0.01	1.00	0.00
4th quintile	0.03	0.13	0.18	0.27	0.32	0.33	0.81	0.27	0.99	0.04	1.00	0.01	1.00	0.00
5th quintile (highest)	0.08	0.20	0.28	0.33	0.43	0.36	0.87	0.23	1.00	0.01	1.00	0.00	1.00	0.00
School type														
Public school	0.02	0.11	0.13	0.24	0.24	0.31	0.71	0.34	0.98	0.08	1.00	0.01	1.00	0.00
Private school	0.06	0.18	0.25	0.31	0.41	0.35	0.85	0.25	1.00	0.02	1.00	0.01	1.00	0.00
¹ Standard deviation. NOTE: Table estimates are based or	tros-sert	ional weights	a within each	VULU Pullor	NO COCMO	C3CW0 C	TCWO CSCV	0 060300	C7CW0) Fe	timates for b	indergarten th	the determinant	t avra de have t	aen nut
on a common scale to support comp	varisons.				·									
SOURCE: U.S. Department of Educ	cation, Natic	onal Center f	or Education	Statistics, Ea	rly Childhoo	d Longitudir	al Study, Kir	ndergarten Cl	ass of 1998-9	99 (ECLS-K)	, fall 1998, s _f	pring 1999, fa	ill 1999, sprii	lg 2000,
spring 2002, spring 2004, and sprin _i	g 2007.							1					1)

	Roun	d 1	Round	d 2	Roune	d 3	Roun	d 4	Rour	d 5	Roun	d 6	Round	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.01	0.08	0.06	0.15	0.12	0.23	0.46	0.33	06.0	0.16	0.96	0.08	96.0	0.03
Sex														
Male	0.01	0.08	0.06	0.15	0.11	0.22	0.43	0.33	0.88	0.18	0.96	0.08	0.98	0.04
Female	0.01	0.08	0.07	0.16	0.14	0.23	0.49	0.32	0.91	0.14	0.97	0.07	0.99	0.03
Race/ethnicity														
White, Non-Hispanic	0.02	0.09	0.07	0.17	0.15	0.25	0.52	0.32	0.93	0.13	0.97	0.06	0.99	0.03
Black, Non-Hispanic	0.01	0.05	0.04	0.11	0.08	0.17	0.35	0.31	0.84	0.20	0.94	0.10	0.97	0.05
Hispanic, race specified	0.01	0.07	0.05	0.13	0.10	0.18	0.41	0.32	0.87	0.18	0.96	0.08	0.98	0.04
Hispanic, race not specified	0.00	0.04	0.03	0.10	0.06	0.14	0.33	0.30	0.82	0.22	0.94	0.10	0.97	0.04
Asian	0.05	0.16	0.13	0.24	0.26	0.33	0.56	0.33	0.92	0.12	0.97	0.08	0.99	0.02
Hawaiian, Other Pacific Islander	0.01	0.07	0.04	0.10	0.07	0.17	0.40	0.32	0.89	0.15	0.96	0.05	0.99	0.02
American Indian, Alaska Native	000	0.02	0.02	0.07	0.03	0.00	0 74	20.77	0.78	0 24	0.01	0.13	0.08	0.04
More than one race, Non-	00.0	70.0	70.0	10.0	c0.0	60.0	+7.0	17.0	00	1.24	16.0	C1.0	0.70	10.0
Hispanic	0.02	0.11	0.07	0.18	0.12	0.21	0.47	0.32	0.91	0.15	0.98	0.05	0.99	0.03
Socioeconomic status														
1st quintile (lowest)	0.00	0.03	0.02	0.06	0.04	0.10	0.27	0.27	0.80	0.23	0.91	0.13	0.96	0.05
2nd quintile	0.01	0.05	0.04	0.12	0.08	0.17	0.40	0.31	0.88	0.17	0.96	0.07	0.98	0.03
3rd quintile	0.01	0.05	0.05	0.12	0.12	0.22	0.47	0.31	0.92	0.13	0.97	0.05	0.99	0.03
4th quintile	0.02	0.08	0.07	0.16	0.15	0.24	0.53	0.31	0.94	0.10	0.98	0.05	0.99	0.02
5th quintile (highest)	0.04	0.13	0.13	0.23	0.23	0.30	0.64	0.30	0.97	0.07	0.99	0.02	1.00	0.01
School type														
Public school	0.01	0.07	0.05	0.14	0.11	0.21	0.44	0.32	0.89	0.17	0.96	0.08	0.98	0.04
Private school	0.03	0.12	0.11	0.21	0.20	0.27	0.60	0.31	0.95	0.08	0.98	0.05	0.99	0.02
¹ Standard deviation.														
NOTE: Table estimates are based of on a common scale to summer form	on cross-sect	ional weight	s within each	round (C1C	W0, C2CW0,	C3CW0, C	4CW0, C5CV	V0, C6CW0,	C7CW0). E	stimates for k	indergarten th	hrough eightl	h grade have l	been put
SOURCE: U.S. Department of Edu	ication. Natio	nal Center f	or Education	Statistics. Ea	ulv Childhoo	d Longitudi	nal Studv. Kir	idergarten Cl	ass of 1998-	99 (ECLS-K)	. fall 1998. sı	nring 1999. f	all 1999. sprii	ng 2000.
spring 2002, spring 2004, and sprin	ng 2007.					0		D				D	J- 4 - 2 - 2 - 100	D D

Probability of proficiency, reading level 5: words in context (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2001-02, 2003–04, 2006–07 Table A20.

	Roun	d 1	Round	12	Round	13	Roun	id 4	Rour	id 5	Roun	id 6	Roun	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.00	0.04	0.01	0.07	0.04	0.12	0.17	0.23	0.67	0.28	0.84	0.20	0.91	0.15
Sex														
Male	0.00	0.04	0.01	0.07	0.03	0.12	0.16	0.23	0.64	0.29	0.83	0.21	0.90	0.16
Female	0.00	0.03	0.01	0.07	0.04	0.13	0.18	0.23	0.70	0.27	0.86	0.19	0.93	0.14
Race/ethnicity														
White, Non-Hispanic	0.00	0.04	0.02	0.08	0.05	0.14	0.20	0.25	0.74	0.25	0.88	0.17	0.95	0.12
Black, Non-Hispanic	0.00	0.02	0.01	0.04	0.02	0.07	0.10	0.17	0.55	0.29	0.76	0.23	0.84	0.19
Hispanic, race specified	0.00	0.04	0.01	0.06	0.02	0.09	0.13	0.20	0.61	0.29	0.81	0.21	06.0	0.17
Hispanic, race not specified	0.00	0.01	0.00	0.04	0.01	0.06	0.09	0.17	0.52	0.30	0.76	0.23	0.86	0.18
Asian	0.01	0.07	0.04	0.12	0.10	0.20	0.25	0.27	0.71	0.25	0.87	0.19	0.95	0.11
Hawaiian, Other Pacific Islander	0.00	0.03	0.01	0.03	0.01	0.05	0.13	0.20	0.60	77.0	0.87	0.18	0 97	012
American Indian, Alaska	00.0	00.0	10.0	00.0	10.0	0.0	CT.0	07.0	0.00	17:0	70.0	01.0	70.0	71.0
Native	0.00	0.00	0.00	0.01	0.00	0.02	0.06	0.13	0.45	0.30	0.69	0.28	0.88	0.17
More than one race, Non- Hispanic	0.01	0.05	0.02	0.00	0.03	010	0.18	0 24	0.68	76.0	0.88	0.16	0 04	0.13
TIINPAILIC	10.0	CO.O	70.0	60.0	c0.0	01.0	01.0	+ 7.0	0.00	17.0	0.00	01.0	+ 6.0	CT-0
Socioeconomic status														
1st quintile (lowest)	0.00	0.01	0.00	0.02	0.01	0.04	0.07	0.13	0.47	0.29	0.69	0.26	0.81	0.21
2nd quintile	0.00	0.02	0.01	0.05	0.02	0.08	0.13	0.19	0.62	0.28	0.81	0.20	06.0	0.15
3rd quintile	0.00	0.01	0.01	0.05	0.03	0.11	0.16	0.21	0.70	0.25	0.87	0.15	0.93	0.12
4th quintile	0.00	0.04	0.02	0.07	0.04	0.13	0.20	0.24	0.76	0.23	06.0	0.15	0.95	0.11
5th quintile (highest)	0.01	0.06	0.04	0.12	0.08	0.19	0.30	0.28	0.84	0.18	0.94	0.09	0.98	0.05
School type														
Public school	0.00	0.03	0.01	0.06	0.03	0.12	0.15	0.22	0.65	0.29	0.83	0.21	0.91	0.16
Private school	0.01	0.06	0.03	0.11	0.06	0.15	0.27	0.27	0.78	0.22	0.91	0.15	0.96	0.10
¹ Standard deviation.														
NOTE: Table estimates are based (on cross-sect.	ional weight.	s within each	round (CIC	W0, C2CW0,	C3CW0, C	34CW0, C5CV	V0, C6CW0,	C7CW0). E	stimates for k	cindergarten tl	hrough eightl	h grade have	been put
SOURCE: U.S. Denartment of Edu	purrovies. Ication. Natic	nal Center f	or Education	Statistics, Ea	rlv Childhood	d Longitudi	nal Study. Kir	ndergarten C	lass of 1998-	99 (ECLS-K)). fall 1998. si	nring 1999. f	all 1999. spri	nø 2000.
spring 2002, spring 2004, and sprir	ng 2007.		10 FULL			יייייקוואד ה	und overy,	uwsurv. ~	>>>> TT 10 0001), 1411 1777 U	, Sund		ш6 4000,

Table A21. Probability of proficiency, reading level 6: literal inference (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000,

	Roun	d 1	Round	12	Roun	d 3	Roun	d 4	Rour	d 5	Roun	d 6	Roune	d 7
Characteristic	Mean	SD^{1}	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.00	0.01	0.00	0.03	0.01	0.06	0.06	0.12	0.42	0.31	0.66	0.30	0.81	0.27
Sex														
Male	0.00	0.01	0.00	0.03	0.01	0.06	0.05	0.12	0.40	0.31	0.64	0.30	0.78	0.29
Female	0.00	0.01	0.00	0.03	0.01	0.06	0.06	0.13	0.45	0.31	0.68	0.29	0.83	0.25
Race/ethnicity														
White, Non-Hispanic	0.00	0.01	0.00	0.04	0.02	0.07	0.07	0.14	0.50	0.31	0.73	0.27	0.87	0.22
Black, Non-Hispanic	0.00	0.01	0.00	0.01	0.00	0.03	0.03	0.08	0.28	0.26	0.52	0.29	0.66	0.31
Hispanic, race specified	0.00	0.02	0.00	0.02	0.01	0.05	0.04	0.10	0.35	0.29	09.0	0.30	0.77	0.29
Hispanic, race not specified	0.00	0.00	0.00	0.02	0.00	0.02	0.03	0.07	0.27	0.26	0.51	0.30	0.70	0.31
Asian	0.00	0.02	0.01	0.05	0.03	0.11	0.09	0.16	0.46	0.30	0.71	0.28	0.88	0.20
Hawaiian, Other Pacific Islander	0.00	0.01	0.00	0.01	0.00	0.01	0.04	0.10	0.33	0.28	09.0	0.29	0.80	0.24
American Indian, Alaska														
Native More than one race Non-	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.21	0.24	0.45	0.34	0.72	0.29
Hispanic	0.00	0.02	0.01	0.04	0.01	0.04	0.06	0.14	0.43	0.30	0.72	0.26	0.86	0.24
Socioeconomic status														
1st quintile (lowest)	0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.05	0.22	0.23	0.43	0.30	0.61	0.32
2nd quintile	00.0	0.01	0.00	0.02	0.00	0.04	0.04	0.09	0.35	0.28	0.59	0.29	0.76	0.27
3rd quintile	0.00	0.00	0.00	0.02	0.01	0.05	0.05	0.10	0.43	0.29	0.69	0.26	0.84	0.23
4th quintile	0.00	0.02	0.00	0.03	0.01	0.06	0.07	0.13	0.51	0.30	0.75	0.25	0.88	0.21
5th quintile (highest)	0.00	0.02	0.01	0.06	0.03	0.10	0.12	0.19	0.64	0.27	0.84	0.19	0.95	0.12
School type														
Public school	0.00	0.01	0.00	0.03	0.01	0.06	0.05	0.12	0.40	0.31	0.64	0.30	0.80	0.27
Private school	0.00	0.02	0.01	0.05	0.02	0.08	0.10	0.16	0.55	0.30	0.77	0.25	0.91	0.18
¹ Standard deviation.														
NOTE: Table estimates are based (on cross-sect.	ional weight.	s within each	round (C1C	W0, C2CW0,	, C3CW0, C	4CW0, C5CV	V0, C6CW0,	C7CW0). E	stimates for k	indergarten tl	hrough eight	n grade have l	been put
SOURCE: U.S. Department of Edu	purrovies. Ication. Natic	nal Center f	or Education (Statistics. Ea	rlv Childhoo	d Longitudii	val Study. Kir	idergarten Cl	ass of 1998-	99 (ECLS-K)	. fall 1998. si	nring 1999. f	all 1999. sprii	ле 2000.
spring 2002, spring 2004, and sprir	ng 2007.		10 דממיימייי	in former of		האיזפוויטים ש	11d1 Divery,	יי ייזאיז איז	~ / / T TA 600		, 1411 J 1/2 1			16 evvv,

Probability of proficiency, reading level 7: extrapolation (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04 and 2006–07 Table A22.

Characteristic Total sample	Round	11	Round	12	Round	13	Roun	d 4	Rour	nd 5	Roun	d 6	Roun	d 7
Total sample	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
	0.00	0.01	0.00	0.02	0.01	0.03	0.03	0.06	0.24	0.21	0.43	0.27	0.64	0.31
Sex														
Male	0.00	0.01	0.00	0.02	0.01	0.03	0.03	0.07	0.22	0.21	0.41	0.27	0.61	0.31
Female	0.00	0.01	0.00	0.02	0.01	0.03	0.03	0.06	0.25	0.21	0.45	0.27	0.67	0.29
Race/ethnicity														
White, Non-Hispanic	0.00	0.01	0.00	0.02	0.01	0.03	0.04	0.07	0.29	0.22	0.50	0.27	0.72	0.27
Black, Non-Hispanic	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.04	0.14	0.15	0.30	0.23	0.45	0.29
Hispanic, race specified	0.00	0.01	0.00	0.01	0.00	0.03	0.02	0.05	0.19	0.19	0.38	0.26	0.59	0.31
Hispanic, race not specified	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.03	0.14	0.15	0.29	0.23	0.50	0.30
Asian Universition Othor Devites	0.00	0.01	0.01	0.02	0.02	0.05	0.05	0.08	0.26	0.21	0.47	0.26	0.74	0.27
Islander Americen Indien Algebo	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.05	0.18	0.17	0.38	0.26	0.61	0.30
Native Native	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.11	0.14	0.28	0.26	0.51	0.29
More than one race, Non-														
Hispanic	0.00	0.01	0.00	0.02	0.00	0.02	0.03	0.08	0.24	0.21	0.48	0.26	0.71	0.28
Socioeconomic status														
1st quintile (lowest)	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.11	0.13	0.24	0.21	0.40	0.29
2nd quintile	0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.04	0.19	0.18	0.36	0.24	0.56	0.30
3rd quintile	0.00	0.00	0.00	0.01	0.00	0.02	0.03	0.05	0.23	0.19	0.44	0.25	0.66	0.28
4th quintile	0.00	0.01	0.00	0.01	0.01	0.03	0.04	0.07	0.29	0.22	0.51	0.25	0.73	0.26
5th quintile (highest)	0.00	0.01	0.01	0.03	0.02	0.05	0.06	0.10	0.39	0.23	0.62	0.24	0.84	0.20
School type														
Public school	0.00	0.01	0.00	0.01	0.01	0.03	0.03	0.06	0.23	0.21	0.41	0.27	0.62	0.31
Private school	0.00	0.01	0.00	0.02	0.01	0.04	0.05	0.08	0.32	0.23	0.55	0.27	0.77	0.25
¹ Standard deviation.							nusu omur			1	17	bala is the second		
TO LE. LAUIC CSUILIAUS ALC DASCU UI		UIIAI WUIGIIIIS			W 0, CZC W 0,	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~		v u, cuc w u,			AILIUU BAI IVII U		I BIAUC IIA VC	occii pui
OII & COTITIOII SCALE TO SUPPOLE COLLIPAL	ITISUIIS.	ant Cantar fo	. Education (Otatiotice For	deildhood	4 Longitudis	ol Study Vin	doncontan C	001 of 1008		\ f~11 1008 c*		-11 1 000 cm	0000

Probability of proficiency, reading level 8: evaluation (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, 2003–04 and 2006–07 Table A23.

spring 2002, spring 2004, and spring 2007.

	Roun	1d 1	Round	12	Roun	d 3	Roun	d 4	Round	15	Round	d 6	Rour	1 T
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.06	0.15	0.27	0.35
šex														
Male	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.06	0.15	0.24	0.33
Female	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.06	0.15	0.30	0.36
Xace/ethnicity														
White, Non-Hispanic	0.00	0.00	0.00	0.00	00.00	0.00	0.00	0.00	0.01	0.05	0.09	0.18	0.35	0.37
Black, Non-Hispanic	0.00	0.00	0.00	0.00	00.0	0.00	0.00	0.00	0.00	0.02	0.02	0.08	0.09	0.21
Hispanic, race specified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.04	0.10	0.21	0.31
Hispanic, race not specified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.08	0.13	0.25
Asian	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.06	0.14	0.40	0.39
Hawanan, Other Facilie Islander American Indian Alastra	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.10	0.26	0.36
Native Native More than one race Non-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.10	0.13	0.24
Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.08	0.18	0.32	0.35
socioeconomic status														
1st quintile (lowest)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.06	0.08	0.19
2nd quintile	00.0	0.00	0.00	0.00	00.0	0.00	0.00	0.00	0.00	0.02	0.03	0.09	0.17	0.28
3rd quintile	00.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.05	0.12	0.24	0.32
4th quintile	00.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.08	0.16	0.35	0.36
5th quintile (highest)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.15	0.23	0.53	0.38
School type														
Public school	00.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.06	0.14	0.25	0.34
Private school	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.12	0.71	0 43	0.39

spring 2002, spring 2004, and spring 2007.

	Roun	d 1	Round	12	Roun	d 3	Roun	id 4	Rour	id 5	Roune	d 6	Roune	17
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.06	0.12
Sex														
Male	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.05	0.11
Female	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.07	0.13
Race/ethnicity														
White, Non-Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.08	0.14
Black, Non-Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	00.0	0.00	0.00	0.00	0.00	0.01	0.02	0.05
Hispanic, race specified	0.00	0.00	0.00	0.00	0.00	0.00	00.0	0.00	0.00	0.00	0.01	0.01	0.04	0.09
Hispanic, race not specified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.07
Asian	0.00	0.00	0.00	0.00	0.00	0.00	00.00	0.00	0.00	0.00	0.01	0.02	0.10	0.15
Hawaiian, Other Pacific														
Islander	0.00	0.00	0.00	0.00	0.00	0.00	00.0	0.00	0.00	0.00	0.01	0.01	0.06	0.11
American Indian, Alaska														
Native	0.00	0.00	0.00	0.00	0.00	0.00	00.0	0.00	0.00	0.00	0.00	0.01	0.02	0.06
More than one race, Non-														
Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.06	0.12
Socioeconomic status														
1st quintile (lowest)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.05
2nd quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07
3rd quintile	0.00	0.00	0.00	0.00	0.00	0.00	00.00	0.00	0.00	0.00	0.01	0.01	0.04	0.09
4th quintile	0.00	0.00	0.00	0.00	0.00	0.00	00.0	0.00	0.00	0.01	0.01	0.02	0.08	0.14
5th quintile (highest)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.14	0.19
School type														
Public school	0.00	0.00	0.00	0.00	0.00	0.00	00.00	0.00	0.00	0.00	0.01	0.02	0.05	0.12
Private school	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.11	0.17
¹ Standard deviation.														
NOTE: Table estimates are based o	in cross-section	ional weights	within each	round (C1C)	W0, C2CW0,	, C3CW0, C	4CW0, C5CV	V0, C6CW0,	C7CW0). E	stimates for k	indergarten th	hrough eightl	h grade have ł	een put
on a common scale to support com	oarisons.				-	:			00010		0001110			
SUUKUE: U.S. Department of Equipsoring 2002, spring 2004, and spring	cation, ivaut g 2007.	onal Center In	or Education	Statistics, Ea	rtly Chilanoo	d Longiuui	nal Study, Nu	ndergarten u	lass of 1990-	99 (ELLA-N)	1, Tall 1998, sf	pring 1999, 1	all 1999, spru	lg zuuu,

	Rour	1 Ju	Roune	12	Roun	d 3	Rour	1d 4	Rour	id 5	Roun	d 6	Roun	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.91	0.18	0.99	0.07	0.99	0.05	1.00	0.02	1.00	0.00	1.00	0.00	1.00	0.00
Sex														
Male	0.91	0.19	0.98	0.07	0.99	0.05	1.00	0.02	1.00	0.00	1.00	0.00	1.00	0.00
Female	0.92	0.17	0.99	0.07	0.99	0.04	1.00	0.02	1.00	0.00	1.00	0.00	1.00	0.00
Race/ethnicity														
White, Non-Hispanic	0.95	0.13	0.99	0.05	1.00	0.03	1.00	0.02	1.00	0.00	1.00	0.00	1.00	0.00
Black, Non-Hispanic	0.88	0.21	0.98	0.09	0.99	0.07	1.00	0.03	1.00	0.00	1.00	0.00	1.00	0.00
Hispanic, race specified	0.87	0.22	0.98	0.08	0.99	0.04	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Hispanic, race not specified	0.81	0.26	0.97	0.10	0.98	0.06	1.00	0.02	1.00	0.00	1.00	0.00	1.00	0.00
Asian	0.96	0.11	1.00	0.03	1.00	0.01	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Hawaiian, Other Pacific														
Islander	0.90	0.21	0.98	0.05	1.00	0.01	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
American Indian, Alaska														
Native	0.81	0.26	0.97	0.08	0.97	0.12	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
More than one race, Non-														
Hispanic	0.93	0.14	0.99	0.06	0.99	0.04	1.00	0.03	1.00	0.00	1.00	0.00	1.00	0.00
Socioeconomic status														
1st quintile (lowest)	0.81	0.26	0.96	0.11	0.98	0.06	1.00	0.03	1.00	0.00	1.00	0.00	1.00	0.00
2nd quintile	0.89	0.20	0.98	0.07	0.99	0.07	1.00	0.03	1.00	0.00	1.00	0.00	1.00	0.00
3rd quintile	0.94	0.15	0.99	0.05	0.99	0.04	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
4th quintile	0.96	0.11	0.99	0.04	1.00	0.02	1.00	0.01	1.00	0.00	1.00	0.00	1.00	0.00
5th quintile (highest)	0.98	0.08	1.00	0.03	1.00	0.00	1.00	0.01	1.00	0.00	1.00	0.00	1.00	0.00
School type														
Public school	0.90	0.19	0.98	0.07	0.99	0.05	1.00	0.02	1.00	0.00	1.00	0.00	1.00	0.00
Private school	0.97	0.10	0.99	0.05	1.00	0.00	1.00	0.02	1.00	0.00	1.00	0.00	1.00	0.00
¹ Standard deviation.	+000 00000 M0	the second s	dooo aidhinn a				LUSU UMUY			timotoo for la	14 and and and the	bda ia damand		4.00 acces
NOLE. LADIE ESUIIIALES ALE DASEU		uollal weight			w 0, L20 w 0,	, UNUUU, U		wu, cucwu,				mougn cigun	I BIAUC HAVE	ocen put
OII a COLIDITE SCALS IN SUPPORT SCAL	Iparisous. Tration Natio	onal Center f	or Education	Statistics Fa	alv Childhoo	d I ongitudi	nal Study Ki	nderaarten C	1908-	00 (FCT S-K)	1 fall 1008 er		11 1 0 00 snri	ол 2000
spring 2002, spring 2004, and sprin	ng 2007.		11 דמתרמויניי	סומווסווס		יייייופווטיד ח	IIdi Diuuy,	י וועטוקמועטו	0111 IN CODI		י יטיריד דושל ין	بد (۲/۲۰۰۶ Jung	ande (2221 III	18 4000,

Probability of proficiency, mathematics level 1: number and shape (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, Table A26.

	Roun	ld 1	Round	12	Round	d 3	Roun	d 4	Roun	id 5	Roune	d 6	Roune	17
Characteristic	Mean	SD^{1}	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.53	0.35	0.83	0.25	0.91	0.19	0.98	60.0	1.00	0.00	1.00	0.00	1.00	0.00
Sex														
Male	0.53	0.36	0.82	0.26	0.90	0.21	0.98	0.10	1.00	0.00	1.00	0.00	1.00	0.00
Female	0.54	0.34	0.84	0.24	0.92	0.17	0.98	0.08	1.00	0.00	1.00	0.00	1.00	0.00
Race/ethnicity														
White, Non-Hispanic	0.63	0.33	0.89	0.20	0.95	0.14	0.99	0.08	1.00	0.00	1.00	0.00	1.00	0.00
Black, Non-Hispanic	0.41	0.33	0.74	0.28	0.86	0.23	0.96	0.12	1.00	0.00	1.00	0.00	1.00	0.00
Hispanic, race specified	0.41	0.34	0.76	0.29	0.88	0.21	0.98	0.09	1.00	0.00	1.00	0.00	1.00	0.00
Hispanic, race not specified	0.32	0.32	0.69	0.31	0.82	0.26	0.98	0.09	1.00	0.00	1.00	0.00	1.00	0.00
Asian	0.66	0.32	0.89	0.18	0.94	0.12	0.99	0.06	1.00	0.00	1.00	0.00	1.00	0.00
Hawaiian, Other Pacific Islander	0.47	0.33	0.78	0.28	0.89	0.16	0.98	0.05	1.00	0.00	1.00	0.00	1.00	0.00
American Indian, Alaska														
Mariye	0.34	0.33	0.73	0.29	0.76	0.30	0.97	0.10	1.00	0.00	1.00	0.00	1.00	0.00
Hispanic	0.53	0.34	0.84	0.23	06.0	0.21	0.98	0.10	1.00	0.00	1.00	0.00	1.00	0.00
Socioeconomic status														
1st quintile (lowest)	030	030	0.68	0 31	0.80	0.27	0 96	013	1 00	0.00	1 00	0.00	1 00	0.00
2nd quintile	0.45	0.34	0.80	0.26	0.89	0.20	0.98	0.11	1.00	0.00	1.00	0.00	1.00	0.00
3rd quintile	0.55	0.33	0.86	0.21	0.94	0.15	0.99	0.08	1.00	0.00	1.00	0.00	1.00	0.00
4th quintile	0.64	0.32	0.90	0.18	0.96	0.12	0.99	0.05	1.00	0.00	1.00	0.00	1.00	0.00
5th quintile (highest)	0.75	0.28	0.94	0.14	0.97	0.08	1.00	0.03	1.00	0.00	1.00	0.00	1.00	0.00
School type														
Public school	0.51	0.35	0.81	0.26	06.0	0.20	0.98	0.09	1.00	0.00	1.00	0.00	1.00	0.00
Private school	0.70	0.30	0.91	0.17	0.98	0.07	1.00	0.05	1.00	0.00	1.00	0.00	1.00	0.00
¹ Standard deviation.														
NOTE: Table estimates are based	on cross-sect	ional weight	s within each	round (C1C	W0, C2CW0,	C3CW0, C	34CW0, C5CV	v0, c6cw0,	C7CW0). Ex	stimates for k	indergarten th	nrough eightl	h grade have l	oeen put
SOURCE: U.S. Department of Edu	ication. Natic	onal Center f	or Education	Statistics. Ea	rlv Childhood	d Longitudi	nal Study. Kin	idergarten Cl	lass of 1998-	99 (ECLS-K)). fall 1998, st	oring 1999. f	all 1999. sprii	ле 2000.
spring 2002, spring 2004, and sprin	ng 2007.					0		0			T (0	I	b 1

Probability of proficiency, mathematics level 2: relative size (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07 Table A27.

	Roun	d 1	Round	12	Roun	d 3	Rour	ıd 4	Rour	id 5	Roun	id 6	Roun	d 7
Characteristic	Mean	\mathbf{SD}^{1}	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.20	0.29	0.53	0.38	0.71	0.35	0.93	0.20	1.00	0.01	1.00	0.00	1.00	0.00
Sex														
Male	0.20	0.30	0.52	0.38	0.70	0.36	0.92	0.20	1.00	0.01	1.00	0.00	1.00	0.00
Female	0.19	0.28	0.53	0.37	0.72	0.34	0.93	0.19	1.00	0.01	1.00	0.00	1.00	0.00
Race/ethnicity														
White, Non-Hispanic	0.26	0.32	0.63	0.35	0.80	0.29	0.96	0.15	1.00	0.01	1.00	0.00	1.00	0.00
Black, Non-Hispanic	0.10	0.19	0.37	0.35	0.59	0.38	0.87	0.26	1.00	0.02	1.00	0.00	1.00	0.00
Hispanic, race specified	0.12	0.23	0.41	0.37	0.65	0.36	0.90	0.23	1.00	0.02	1.00	0.00	1.00	0.00
Hispanic, race not specified	0.08	0.18	0.33	0.35	0.52	0.39	0.89	0.22	1.00	0.02	1.00	0.00	1.00	0.00
Asian	0.31	0.35	0.62	0.36	0.76	0.33	0.94	0.17	1.00	0.01	1.00	0.00	1.00	0.00
Hawaiian, Other Pacific														
Islander	0.13	0.24	0.41	0.35	0.57	0.35	0.90	0.20	1.00	0.02	1.00	0.00	1.00	0.00
American Indian, Alaska														
Native	0.09	0.19	0.36	0.35	0.44	0.38	0.88	0.23	1.00	0.01	1.00	0.00	1.00	0.00
More than one race, Non-														
Hispanic	0.19	0.29	0.51	0.36	0.72	0.34	0.93	0.19	1.00	0.01	1.00	0.00	1.00	0.00
-														
Socioeconomic status														
1st quintile (lowest)	0.06	0.15	0.29	0.32	0.47	0.38	0.85	0.27	1.00	0.02	1.00	0.00	1.00	0.00
2nd quintile	0.13	0.23	0.46	0.37	0.65	0.35	0.91	0.21	1.00	0.01	1.00	0.00	1.00	0.00
3rd quintile	0.18	0.27	0.55	0.36	0.76	0.30	0.95	0.17	1.00	0.01	1.00	0.00	1.00	0.00
4th quintile	0.25	0.31	0.63	0.35	0.82	0.27	0.96	0.13	1.00	0.01	1.00	0.00	1.00	0.00
5th quintile (highest)	0.39	0.36	0.74	0.31	0.88	0.23	0.98	0.09	1.00	0.00	1.00	0.00	1.00	0.00
School type														
Public school	0.18	0.28	0.50	0.38	0.69	0.35	0.92	0.20	1.00	0.01	1.00	0.00	1.00	0.00
Private school	0.33	0.34	0.68	0.34	0.87	0.22	0.98	0.10	1.00	0.00	1.00	0.00	1.00	0.00
¹ Standard deviation.														
NOTE: Table estimates are based of	on cross-secti	onal weights	s within each	round (C1C)	W0, C2CW0,	C3CW0, C	4CW0, C5CV	V0, C6CW0,	C7CW0). E	stimates for k	cindergarten tj	hrough eight	h grade have	been put
on a common scale to support com	parisons.													
SOURCE: U.S. Department of Edu	ucation, Natic	nal Center t	or Education	Statistics, Ea	urly Childhoo	d Longitudi	nal Study, Ki	ndergarten C	lass of 1998-	99 (ECLS-K), fall 1998, s	pring 1999, 1	all 1999, spri	ng 2000,
spring 2002, spring 2004, and sprin	ng 2007.													

Probability of proficiency, mathematics level 3: ordinality, sequence (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07 Table A28.

1	Roun	d 1	Roune	12	Roun	d 3	Roun	id 4	Rour	id 5	Roun	9 P	Round	17
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.03	0.11	0.16	0.25	0.31	0.33	0.69	0.33	96.0	0.11	0.99	0.03	1.00	0.00
Sex														
Male	0.04	0.13	0.16	0.26	0.32	0.34	0.69	0.34	0.96	0.11	0.99	0.03	1.00	0.00
Female	0.03	0.10	0.15	0.24	0.30	0.32	0.70	0.32	0.96	0.10	0.99	0.03	1.00	0.00
Race/ethnicity														
White, Non-Hispanic	0.05	0.14	0.21	0.28	0.38	0.34	0.77	0.29	0.98	0.08	1.00	0.02	1.00	0.00
Black, Non-Hispanic	0.01	0.05	0.07	0.16	0.20	0.27	0.55	0.35	0.93	0.14	0.99	0.04	1.00	0.01
Hispanic, race specified	0.01	0.07	0.10	0.19	0.25	0.30	0.63	0.35	0.95	0.13	0.99	0.03	1.00	0.00
Hispanic, race not specified	0.01	0.04	0.07	0.16	0.16	0.24	0.57	0.34	0.94	0.13	0.99	0.03	1.00	0.00
Asian	0.08	0.19	0.23	0.30	0.40	0.36	0.72	0.33	0.97	0.08	0.99	0.03	1.00	0.00
Hawaiian, Other Pacific														
Islander	0.02	0.10	0.09	0.18	0.16	0.24	0.56	0.34	0.95	0.13	1.00	0.01	1.00	0.00
American Indian, Alaska														
Native	0.01	0.05	0.07	0.16	0.12	0.21	0.52	0.35	0.93	0.13	0.99	0.04	1.00	0.01
More than one race, Non-														
Hispanic	0.03	0.11	0.14	0.23	0.26	0.29	0.69	0.33	0.96	0.11	0.99	0.03	1.00	0.00
Socioeconomic status														
1st quintile (lowest)	0.00	0.04	0.05	0.13	0.14	0.24	0.51	0.35	0.92	0.16	0.98	0.05	1.00	0.01
2nd quintile	0.01	0.06	0.11	0.20	0.22	0.28	0.64	0.34	0.96	0.11	0.99	0.03	1.00	0.00
3rd quintile	0.02	0.08	0.14	0.22	0.31	0.31	0.72	0.31	0.97	0.08	1.00	0.02	1.00	0.00
4th quintile	0.04	0.12	0.20	0.26	0.37	0.32	0.78	0.28	0.98	0.07	1.00	0.01	1.00	0.00
5th quintile (highest)	0.09	0.19	0.30	0.32	0.52	0.36	0.86	0.23	0.99	0.03	1.00	0.01	1.00	0.00
School type														
Public school	0.03	0.10	0.14	0.23	0.29	0.32	0.68	0.34	0.96	0.11	0.99	0.03	1.00	0.00
Private school	0.07	0.17	0.25	0.30	0.45	0.33	0.81	0.25	0.98	0.05	1.00	0.01	1.00	0.00
¹ Standard deviation.														
NOTE: Table estimates are based or	1 cross-secti	ional weights	within each	round (C1C)	W0, C2CW0,	, C3CW0, C	4CW0, C5CV	V0, C6CW0,	C7CW0). E.	stimates for k	indergarten tl	hrough eight	h grade have l	been put
on a common scale to support comp	arisons.	1												
SOURCE: U.S. Department of Educ suring 2002, spring 2004, and spring	cation, Nauc r 2007.	onal Center 10	or Education	Statistics, Ea	rly Chiidnoo	d Longituai	nal Study, Kıı	ndergarten U	lass of 1998-	99 (ECLS-K), tall 1998, s _l	pring 1999, 1	iall 1999, spri	ıg 2000,

	Roun	d 1	Rounc	12	Round	d 3	Roun	ıd 4	Rour	ıd 5	Roun	d 6	Roune	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.00	0.03	0.01	0.07	0.04	0.14	0.22	0.29	0.75	0.32	0.92	0.20	0.98	0.08
Sex														
Male	0.00	0.04	0.02	0.09	0.05	0.16	0.24	0.31	0.76	0.32	0.92	0.19	0.98	0.09
Female	0.00	0.01	0.01	0.05	0.04	0.12	0.20	0.27	0.74	0.32	0.91	0.20	0.98	0.07
Race/ethnicity														
White, Non-Hispanic	0.00	0.04	0.02	0.09	0.06	0.16	0.29	0.32	0.83	0.27	0.95	0.16	0.99	0.06
Black, Non-Hispanic	0.00	0.02	0.00	0.03	0.01	0.07	0.09	0.18	0.57	0.36	0.84	0.27	0.95	0.12
Hispanic, race specified	0.00	0.02	0.01	0.04	0.02	0.08	0.16	0.25	0.68	0.34	0.91	0.20	0.98	0.07
Hispanic, race not specified	0.00	0.00	0.00	0.03	0.01	0.03	0.10	0.19	0.64	0.35	0.88	0.23	0.97	0.10
Asian	0.01	0.05	0.03	0.13	0.09	0.22	0.27	0.32	0.79	0.31	0.95	0.17	0.99	0.06
Hawaiian, Other Pacific														
Islander	0.00	0.00	0.01	0.05	0.01	0.04	0.10	0.19	0.70	0.33	0.93	0.15	0.98	0.07
American Indian, Alaska														
Native	0.00	0.00	0.00	0.04	0.01	0.05	0.09	0.17	0.56	0.36	0.79	0.28	0.97	0.09
More than one race, Non-														
Hispanic	0.00	0.04	0.01	0.07	0.03	0.11	0.23	0.30	0.77	0.31	0.93	0.19	0.99	0.05
Socioeconomic status														
1st quintile (lowest)	0.00	0.01	0.00	0.02	0.01	0.04	0.08	0.17	0.55	0.36	0.80	0.29	0.95	0.12
2nd quintile	0.00	0.00	0.01	0.04	0.02	0.10	0.15	0.24	0.70	0.33	0.91	0.20	0.97	0.09
3rd quintile	0.00	0.01	0.01	0.06	0.03	0.10	0.21	0.27	0.78	0.29	0.95	0.14	0.99	0.06
4th quintile	0.00	0.03	0.02	0.08	0.05	0.13	0.28	0.30	0.85	0.25	0.96	0.14	0.99	0.05
5th quintile (highest)	0.01	0.06	0.04	0.13	0.12	0.23	0.41	0.35	0.91	0.19	0.98	0.09	1.00	0.03
School type														
Public school	0.00	0.03	0.01	0.07	0.04	0.13	0.21	0.28	0.74	0.33	0.91	0.20	0.98	0.08
Private school	0.01	0.05	0.03	0.11	0.08	0.19	0.32	0.33	0.83	0.26	0.96	0.14	0.99	0.05
¹ Standard deviation.														
NOTE: Table estimates are based o	in cross-sect	ional weights	s within each	round (C1C)	W0, C2CW0,	C3CW0, C	4CW0, C5CV	V0, C6CW0,	C7CW0). E	stimates for k	cindergarten tl	hrough eightl	h grade have l	been put
on a common scale to support component component component of the providence of the	barisons.	for the former	an Ddunottion (Statistics Do	alse Childhao	d I amainth	and Chindre Via	o representer	1000 of 1000		- £11 1000	7 000		0000
SUUKUE: U.S. Department of Louis spring 2002, spring 2004, and spring	cauon, mau g 2007.	JIIAI UGIIKA I	OF Education	DIAUISUUS, LIC	urly cumuo	d Luigiuur	nai ouuy, m	ndergauen –	-0661 10 1220-	N-6773) 66), 1411 1770, J	priug 1777, 1	all 1 <i>777</i> , spin	ng zuw,

Probability of proficiency, mathematics level 5: multiplication/division (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000. 2001-02. 2003–04. and 2006–07 Table A30.

Characteristic Mean Total sample 0.00	tound 1	Roun	d 2	Roun	d 3	Roun	id 4	Rour	d 5	Roun	d 6	Roun	d 7
Total sample 0.00	n SD ¹	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
C	0 0.01	0.00	0.02	0.00	0.03	0.03	0.11	0.41	0.40	0.72	0.37	0.89	0.25
2CX													
Male 0.00	0 0.01	0.00	0.02	0.00	0.04	0.04	0.13	0.44	0.40	0.75	0.36	0.88	0.26
Female 0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.09	0.37	0.38	0.69	0.38	0.89	0.25
Race/ethnicity													
White, Non-Hispanic 0.00	0 0.00	0.00	0.02	0.00	0.04	0.04	0.13	0.50	0.40	0.80	0.33	0.93	0.21
Black, Non-Hispanic 0.00	00.0 0	0.00	0.01	0.00	0.01	0.01	0.04	0.20	0.31	0.53	0.40	0.76	0.33
Hispanic, race specified 0.00	00.00	0.00	0.00	0.00	0.01	0.02	0.08	0.32	0.37	0.67	0.38	0.89	0.24
Hispanic, race not specified 0.00	00.00	0.00	0.00	0.00	0.00	0.01	0.04	0.25	0.34	0.63	0.40	0.84	0.29
Asian 0.00	00.00	0.00	0.04	0.01	0.06	0.05	0.17	0.52	0.42	0.83	0.32	0.93	0.22
Hawaiian, Other Pacific													
Islander 0.00	00.00	0.00	0.00	0.00	0.00	0.00	0.03	0.28	0.32	0.68	0.37	0.91	0.23
American Indian, Alaska													
Native 0.00	00.00	0.00	0.00	0.00	0.00	0.01	0.06	0.20	0.32	0.49	0.42	0.82	0.31
More than one race, Non-													
Hispanic 0.00	0 0.03	0.00	0.03	0.00	0.01	0.02	0.08	0.43	0.39	0.73	0.35	0.93	0.21
Socioeconomic status													
1et mintile (lowest)	000	000	000	000	000	0.01	0.05	0.18	0.00	0.47	0.41	0.76	0 34
and ministile (10 more) 0.00		00.0	0.00	0000	0.00	10.0	90.0	0.21	036	0.66	0.39	0.04	000
	00.0	0.00	0.00	00.0	20.0	10.0	00.0	10.0	00.0			10.0	(7.0 100
	0.00	0.00	0.01	0.00	0.01	0.02	60.0	0.41	0.38	0./0	0.33	76.0	17.0
4th quintile 0.00	0.00	0.00	0.02	0.00	0.02	0.03	0.11	0.53	0.39	0.83	0.30	0.95	0.17
5th quintile (highest) 0.00	0 0.01	0.00	0.03	0.01	0.06	0.08	0.18	0.66	0.37	0.91	0.22	0.98	0.11
School type													
Public school 0.00	0 0.01	0.00	0.01	0.00	0.03	0.03	0.11	0.40	0.39	0.71	0.38	0.88	0.26
Private school 0.00	0 0.01	0.00	0.03	0.01	0.04	0.05	0.14	0.50	0.40	0.83	0.29	0.96	0.16
¹ Standard deviation.	•												
NOTE: Table estimates are based on cross-	sectional weight	ghts within each	round (CIC	wu, czcwu,	C3CW0, C	4CW0, C5CV	vu, c6cwu,	C7CW0). E	stimates for k	indergarten tl	hrough eightl	h grade have	been put
on a common scale to support comparisons.	Viational Canto	- for Education	Ctatiotice Ea	oodhlid Other	d I anditudia	"-1 Chida Vir	O networker	1 م4 1002	ייי יבינו א וגי	11 1008 c		-11 1000 con	0000
SUUNCE: U.S. DEPARTIMENT OF EQUATION, 17 Spring 2002, Spring 2004, and Spring 2007.	Nauonai John	I I I Euucanon	DIAUSUUS, LIA	TIY CIIIUUW	d Lungiuuu	lal ouuy, nu	ndergarten 🗸	1855 UL 1770-	49 (EULID-IN)), 1all 1770, o	priug 1777, 1	all 1 <i>77</i> 7, spu	ng zuw,

	Roun	d 1	Round	12	Roun	d 3	Rour	ld 4	Rour	id 5	Roun	d 6	Roun	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	0.00	0.00	0.00	00.00	0.00	0.00	0.00	0.02	0.13	0.23	0.40	0.37	0.66	0.37
Sex														
Male	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.15	0.25	0.44	0.38	0.67	0.37
Female	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.11	0.20	0.37	0.37	0.64	0.37
Race/ethnicity														
White, Non-Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.17	0.25	0.50	0.38	0.74	0.33
Black, Non-Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.14	0.19	0.27	0.42	0.37
Hispanic, race specified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.09	0.19	0.33	0.35	0.62	0.37
Hispanic, race not specified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.15	0.29	0.33	0.55	0.38
Asian	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.20	0.28	0.56	0.39	0.78	0.33
Hawaiian, Other Pacific														
Islander	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.15	0.37	0.37	0.68	0.37
American Indian, Alaska														
Native	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.11	0.20	0.30	0.54	0.39
More than one race, Non-														
Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.13	0.23	0.42	0.39	0.75	0.34
Socioaconomio etatue														
	0	0	0	0	0	0	0							
Ist quintile (lowest)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.11	0.19	0.28	0.43	0.38
2nd quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.08	0.16	0.29	0.32	0.56	0.38
3rd quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.11	0.19	0.40	0.36	0.69	0.34
4th quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.18	0.25	0.52	0.36	0.77	0.31
5th quintile (highest)	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.28	0.31	0.67	0.35	0.87	0.24
School type														
Public school	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.12	0.22	0.39	0.37	0.64	0.38
Private school	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.17	0.26	0.52	0.37	0.78	0.30
¹ Standard deviation.														
NOTE: Table estimates are based o	in cross-sect	ional weight:	s within each	round (C1C)	W0, C2CW0,	C3CW0, C	4CW0, C5CV	V0, C6CW0,	C7CW0). E	stimates for k	cindergarten tl	hrough eight	h grade have	been put
on a common scale to support com	parisons.	1 Canton 6	Ddunotion	Ototiotion Do	Childhao	ibudiano 1 t	1 0 t D	Jucture -	1 م1000		· 5-11 1000	1000 4	-11 1 0 00 2001	
SUUKUE: U.S. Department of Equal spring 2002, spring 2004, and spring	g 2007.	onal Center 1	of Education	otausucs, da	riy Cillianoo	d Louguuu	nal ətudy, nu	Juergarien ∽	1855 UI 1770-	79 (EULD-D), Iàu 1770, 5J	prilig 1777, 1	all 1777, spil	ng zuuu,

Characteristic M Total sample (Round	- 	Round	12	Round	13	Roun	d 4	Rour	id 5	Roun	d 6	Roun	d 7
Total sample	dean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.12	0.25	0.36	0.41
Sex														
Male	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.14	0.27	0.38	0.41
Female	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.10	0.22	0.33	0.40
Race/ethnicity														
White, Non-Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	00.0	0.00	0.02	0.08	0.16	0.28	0.44	0.42
Black, Non-Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.02	0.11	0.12	0.26
Hispanic, race specified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.08	0.19	0.29	0.37
Hispanic, race not specified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.17	0.25	0.36
Asian $\Delta t_{1} = \Delta t_{1} = \Delta t_{2}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.24	0.34	0.56	0.44
hawanan, Omer Facinc Islander	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.20	0.38	0.39
American Indian, Alaska))	5		•			•) 		
Native	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.15	0.21	0.34
More than one race, Non-			00 0	00 0		000	000	000						, ,
Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	c0.0	0.16	67.0	0.47	0.43
Socioeconomic status														
1st quintile (lowest)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.13	0.15	0.29
2nd quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.16	0.24	0.35
3rd quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.09	0.21	0.34	0.38
4th quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.15	0.27	0.47	0.41
5th quintile (highest)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.12	0.29	0.34	0.64	0.40
School type														
Public school	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.11	0.24	0.34	0.40
Private school (0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.17	0.29	0.47	0.42
¹ Standard deviation.		-	-				ILUSU VIILUP			- -	1 1 1		-	
	USS-Section			נסחוום (כוכי	vu, uzuwu,	C) (NA) (C)	40 WU, UJU V	vu, cocwu,	C/CWU). E	sumates tot k	unuergarien u	mougn eigim	I graue Have	neen put
OR a COMPLION SCARE tO SUPPORT COMPARISO COMPACE: 11 C. Demartment of Education	ons. • Nations	al Conter for	. Education S	Statistics Far	4. Childhood	1 T anaitudir	ol Study Kin	dargartan Cl	2001 Jone	ייי יברו פ עו	£-11 1000			000c ~~

spring 2002, spring 2004, and spring 2007.

Probability of proficiency, mathematics level 8: fractions (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2003–04, and 2006–07

Table A33.

	Roun	1 Ju	Round	12	Roun	d 3	Rour	ld 4	Rour	ld 5	Roun	id 6	Roun	d 7
Characteristic	Mean	SD^1	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total sample	00.0	0.00	0.00	00.00	0.00	0.00	0.00	0.00	00.00	0.02	0.02	0.09	0.15	0.29
Sex														
Male	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.11	0.17	0.30
Female	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.07	0.14	0.28
Race/ethnicity														
White, Non-Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.10	0.20	0.32
Black, Non-Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.03	0.14
Hispanic, race specified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.06	0.10	0.23
Hispanic, race not specified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.08	0.21
Asian	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.15	0.35	0.41
Hawaiian, Other Pacific														
Islander	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.11	0.22
American Indian, Alaska														
Native	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.08	0.23
More than one race, Non-														
Hispanic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.24	0.35
Socioeconomic status														
1st mintle (langet)	000	000	000	000	000	000	000	000	000	000	000	0.02	10.0	0.15
	00.0	00.0	00.0	0.00	00.0	00.0	00.0	00.0	00.0	00.0	0.00	50.0	+ 0.0	CT-0
Znd quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	c0.0	0.08	0.21
3rd quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.05	0.13	0.26
4th quintile	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.08	0.20	0.31
5th quintile (highest)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.15	0.35	0.38
School type														
Public school	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.09	0.15	0.29
Private school	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.22	0.33
¹ Standard deviation.			114:					0/11/2/2/0/1		1		1	1	
NOLE: Lable estimates are based (UII CIUSS-SECI	uollal weiglil.			wu, uzu wu,	n non co	40 MO, COC V	പം, പാപം,	C/CWO). E	sumates for k	unuergarien u	mougn cigin	II graue nave	neen put
OII a COILILIOLI SCALE LO SUPPOLI COLLI COLTD/F+11 C Densitment of Edu	ipation Natio	onal Cantar f	or Education '	Ctatictice Ha	-dv Childhoo	d I onaitudi	nol Study Ki	Jaraarten C	امدد مf 1008_	00 (ECT S-K	o fall 1008 e		11 1 000 enri	
SUCINCE: U.S. Department of Low spring 2002, spring 2004, and sprin	1g 2007.	Ultal Currer 1	UL EUUVAUVI	olausuvs, ra		ת בטווצווטע	uzi (unic igi	Jucigaricu 🗸	1455 UL 1220-	או-טישרש) לל), Idll 1770, o	pung 1777, 1	unde <i>1222</i> , ann	IIB ZUUU,

Probability of proficiency, mathematics level 9: area and volume (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001-02, 2001-02, 2003–04, and 2006–07 Table A34.

	Round 1 Modal Level=1	Round 2 Modal Level=3	Round 3 Modal Level=3	Round 4 Modal Level=4	Round 5 Modal Level=6	Round 6 Modal Level=7	Round 7 Modal Level=8
Characteristic	Percent						
Total sample	64.50	48.55	65.19	77.63	71.44	70.39	63.39
Sex							
Male	60.73	45.19	60.48	73.44	68.34	68.27	59.83
Female	68.50	52.13	70.15	82.08	74.80	72.64	67.24
Race/ethnicity							
White, Non-Hispanic	69.84	55.02	72.49	83.05	80.31	79.12	71.89
Black, Non-Hispanic	59.90	33.89	51.69	67.61	58.88	55.46	46.80
Hispanic, race specified	51.92	44.32	63.01	73.87	63.87	61.58	57.75
Hispanic, race not specified	46.73	38.00	50.31	65.04	52.03	52.12	47.64
Asian	81.53	63.44	68.95	84.86	73.60	76.13	69.50
Hawaiian, Other Pacific Islander	63.99	35.16	42.05	75.78	61.59	62.61	56.59
American Indian, Alaska Native	34.68	26.82	24.84	49.99	43.90	43.09	48.23
More than one race, Non-Hispanic	62.72	44.63	65.28	82.27	72.67	81.37	67.73
socioeconomic status							
1st quintile (lowest)	41.72	26.99	40.07	59.00	46.95	42.50	40.19
2nd quintile	56.62	39.13	55.71	73.79	65.84	63.48	55.13
3rd quintile	64.37	48.27	67.70	80.24	76.13	76.98	64.19
4th quintile	73.72	58.29	76.52	85.83	82.20	81.24	73.34
5th quintile (highest)	83.85	70.11	83.37	90.56	90.77	91.72	85.05
School type							
Public school	61.16	45.33	63.06	75.87	69.73	68.73	62.00
Private school	82,79	66.15	83 43	8015	8516	83.66	76.07

Table A35. Percent of children at or above modal reading proficiency for each grade: School years 1998–99, 1999–2000, 2001–02,

grade have been put on a common scale to support comparisons. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring, 2004, and spring 2007.

Characteristic Pacent Pacent <th< th=""><th></th><th>Round 1 Modal Level=1</th><th>Round 2 Modal Level=3</th><th>Round 3 Modal Level=3</th><th>Round 4 Modal Level=4</th><th>Round 5 Modal Level=5</th><th>Round 6 Modal Level=6</th><th>Round 7 Modal Level=7</th></th<>		Round 1 Modal Level=1	Round 2 Modal Level=3	Round 3 Modal Level=3	Round 4 Modal Level=4	Round 5 Modal Level=5	Round 6 Modal Level=6	Round 7 Modal Level=7
	Characteristic	Percent						
Sax Male 2.281 5.2.61 6.9.83 7.0.17 7.1.45 6.7.48 Female 5.7.98 5.4.20 7.1.58 70.17 7.2.79 6.8.57 6.1.60 Female 5.7.98 5.4.20 7.1.58 70.17 7.2.79 6.8.57 6.1.60 Racefunicity 6.6.6 6.485 8.0.44 7.8.05 8.2.5.4 81.06 7.3.8 While, Non-Hispanic 6.6.6 6.485 8.0.44 7.8.05 8.2.5.4 81.06 7.3.8 Hispanic, near or specified 4.3.77 3.8.8 0.44 7.8.05 8.2.5.8 6.4.1 6.4.1 8.4.4 Hispanic, near or specified 4.3.77 3.8.101 4.6.3.6 5.5.68 6.6.41 6.4.1 8.4.4 Hispanic, near or specified 2.7.67 3.0.11 4.6.3.6 5.5.68 6.6.41 6.4.4 3.3.7 Asim Anariom Induce Pacifie Induce 5.7.38 5.1.9 17.3.8 84.46 7.7.41 Hawin, Ohrer Pacifie Induce 5.7.9 3.0.13 7.7.3 8.4.6 7.7.43 8.4.6 7.7.41 Anariom Induce Pacifie Induce 5.7.9 3.0.13 7.7.3 8.4.6 7.7.3 8.4.6 7.7.41 Anariom Induce Pacifie Induce 5.7.9 3.0.13 7.7.3 8.4.6 7.7.3 8.4.6 7.7.41 Anariom Induce Pacifie Induce 5.7.9 5.7.9 8.0.47 7.1.97 7.3.4 8.4.6 7.7.41 Anariom Induce Pacifie Induce 5.7.9 4.2.1 7.7.3 8.6.4 4.4.6 4.4.5 6.4.1 7.7.3 8.6.4 4.4.6 4.4.5 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.4 4.7.5 7.7.3 8.6.4 4.4.6 4.7.5 7.3.4.6 4.4.6 4.7.5 7.3.4.6 4.4.6 4.7.5 7.3.4.6 4.4.6 4.7.5 7.3.4.6 4.4.6 6.6.3 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.6.8 5.7.7 7.3.9 8.6.7 5.7.3 8.6.1 7.3.4 8.2.1 7.3.4 8.2.2 5.4.8 10.10 8.5.4 7.7.3 8.6.1 7.3.4 8.2.2 5.4.8 10.10 8.5.4 7.7.3 8.6.1 7.3.4 8.2.5 8.6.7 5.7.4 7.3.4 8.2.2 5.4.8 10.10 8.5.4 7.7.3 8.6.1 7.3.4 8.2.5 8.6.7 7.3.4 8.2.5 8.6.7 7.3.4 8.2.5 8.6.7 7.3.4 8.2.5 8.6.7 7.3.4 8.2.5 8.6.7 7.3.4 8.2.5 8.6.7 7.3.4 8.2.5 8.6.7 7.3.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.4 8.2.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5 8.6.7 7.5.5	Total sample	55.34	53.38	70.69	70.16	74.29	73.15	64.72
Male 52.81 52.61 69.83 70.15 75.71 77.45 67.48 Fenule 57.98 54.20 71.58 70.17 72.79 66.57 61.69 Raceelmicity 66.06 64.85 80.44 78.05 85.54 81.06 73.48 White, Non-Hispanic 66.06 64.85 80.44 78.05 85.54 81.06 73.48 Black, Non-Hispanic 66.06 64.85 80.44 78.05 85.54 87.00 73.48 Hispanic, race or specified 10.3 30.11 46.36 56.40 57.38 57.38 57.34 50.40 Asim 09.46 61.45 72.25 70.43 77.38 84.46 77.41 Asim 09.46 61.45 72.25 70.43 77.38 84.46 77.41 Asim 09.46 61.4 47.21 77.38 84.46 77.41 Asim 09.46 61.4 72.25 70.43 77.38 84.46 <td>Sex</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>	Sex							
Female 57.38 54.20 71.58 70.17 72.79 68.57 61.60 Racechnicity Mine, Non-Hispanic 66.06 64.85 80.44 78.05 83.54 81.06 73.83 Whie, Non-Hispanic 66.06 64.85 80.44 78.05 83.56 64.10 73.14 Hispanic, race specified 40.27 38.63 01.51 62.22 68.20 67.14 53.44 Hispanic, race specified 40.27 38.63 01.51 62.22 68.20 67.14 53.43 Asim 69.46 61.41 72.25 77.38 84.46 77.31 Asim 69.46 61.47 72.25 70.43 77.38 84.46 77.31 Asim 69.46 61.47 72.25 70.43 77.34 77.34 Asim 69.47 31.17 71.24 48.21 77.34 77.34 Asim 69.46 51.91 70.75 69.56 80.47 71.97 77.34	Male	52.81	52.61	69.83	70.15	75.71	77.45	67.48
Racefehnicity Racefehnicity Racefehnicity 73.83 81.06 73.83 81.06 73.83 81.06 73.83 83.44 78.05 82.54 81.06 73.83 83.44 73.05 82.53 63.70 42.10 73.83 Hispanic, race specified 40.27 38.63 61.51 62.22 68.20 67.14 58.44 53.70 42.10 Hispanic, race specified 40.27 38.63 61.51 62.22 68.20 67.14 58.44 53.73 53.70 42.10 Asian 60.46 61.45 51.91 46.36 55.68 66.41 64.64 53.73 53.73 53.73 53.73 53.73 53.73 53.73 53.70 53.73 53.70 53.70 53.74 53.73 53.70 53.70 53.70 53.70 53.70 53.70 53.70 53.70 53.70 53.70 53.70 53.70 53.70 53.70 53.70	Female	57.98	54.20	71.58	70.17	72.79	68.57	61.69
White, Non-Hispanic66.0664.8580.4478.0582.5481.0673.83Black, Non-HispanicH3737.2439.9757.8053.5853.7073.14Hispanic, race specified 2.77 30.1161.5162.2268.2067.1453.44Hispanic, race specified 2.77 30.1161.4572.2568.2067.1453.37Asim 69.46 61.45 30.11 24.26 50.4054.0677.3854.33AsimOther Pacific Islander53.36 39.18 56.40 54.06 70.7082.5069.70American Indian, Alaska Native 34.97 34.17 41.24 48.21 56.89 54.34 53.38American Indian, Alaska Native 34.77 41.24 48.21 56.89 54.34 53.38American Indian, Alaska Native 34.97 34.17 41.24 48.21 56.89 54.34 53.33 More than one race, Non-Hispanic 57.78 54.16 77.78 77.34 77.34 Socioeconomic status 27.77 41.24 42.41 49.97 56.89 54.34 53.36 Jad quintile (lowest) 28.62 27.94 47.99 56.96 67.32 67.32 54.86 Jad quintile (lowest) 78.70 67.37 79.41 77.91 77.94 67.32 67.32 Jad quintile (lowest) 77.96 88.31 86.79 90.42 77.99 87.36 67.32	Race/ethnicity							
Black, Non-Hispanic 4.37 37.24 59.97 57.80 53.58 53.70 42.10 Hispanic, race specified 40.27 38.63 61.51 62.22 68.20 67.14 58.43 Hispanic, race specified 40.27 38.63 61.45 72.25 70.43 77.41 53.37 Hispanic, race not specified 27.67 30.11 46.56 55.68 66.71.4 58.44 57.31 Hawian, Other Pacific Islader 53.36 61.45 72.25 70.43 77.41 58.44 57.34 53.35 American India, Alaska Naive 34.77 41.24 48.21 66.89 54.34 53.35 More than one race, Non-Hispanic 57.39 34.17 41.24 48.21 70.43 71.97 73.45 More than one race, Non-Hispanic 57.39 54.46 77.41 77.41 More than one race, Non-Hispanic 57.39 54.34 70.35 54.34 73.45 Socioeconomic status Sacio 57.30 57.34 <th< td=""><td>White, Non-Hispanic</td><td>66.06</td><td>64.85</td><td>80.44</td><td>78.05</td><td>82.54</td><td>81.06</td><td>73.83</td></th<>	White, Non-Hispanic	66.06	64.85	80.44	78.05	82.54	81.06	73.83
Hispanic, race specified 40.27 38.63 61.51 62.22 68.20 67.14 58.44 Hispanic, race specified 27.67 30.11 46.36 55.68 66.41 64.64 53.79 Asian 69.46 61.45 30.25 53.68 66.41 64.64 53.79 Asian 69.46 61.45 52.64 55.68 66.41 64.64 53.79 American Indian, Other Pacific Islander 33.36 39.18 56.40 54.06 70.70 82.50 69.70 American Indian, Other Pacific Islander 57.38 31.17 41.24 48.21 56.89 54.34 53.38 American Indian, Alaska Native 57.78 51.91 70.72 69.56 80.47 71.97 73.45 American Indian, Alaska Native 57.78 51.91 70.72 69.56 80.47 71.97 73.45 American Indian, Alaska Native 57.78 51.91 70.72 69.56 80.47 71.97 73.45 American Indian, Alaska Native 57.79 57.96 57.04 42.41 49.97 77.53 42.23 Socioeconomic status 147.09 76.74 81.47 72.71 77.99 77.68 87.26 Socioeconomic status 77.96 57.96 57.96 57.96 77.04 87.96 77.53 42.23 Socioeconomic status 77.97 72.71 77.97 77.96 77.68 87.86 Socioeconomic status </td <td>Black, Non-Hispanic</td> <td>44.37</td> <td>37.24</td> <td>59.97</td> <td>57.80</td> <td>53.58</td> <td>53.70</td> <td>42.10</td>	Black, Non-Hispanic	44.37	37.24	59.97	57.80	53.58	53.70	42.10
Hispanic race not specified 27.67 30.11 46.36 55.68 66.41 64.64 53.79 Asian 09.46 61.45 72.25 70.43 77.38 84.46 77.41 Hawiian, Other Pacific Islander 53.36 39.18 56.40 54.06 70.70 82.50 69.70 American Indian, Alaska Native 34.97 34.17 41.24 48.21 56.89 54.34 53.38 American Indian, Alaska Native 34.17 41.24 48.21 56.89 54.34 53.33 American Indian, Alaska Native 34.17 41.24 48.21 56.89 54.34 53.33 American Indian, Alaska Native 34.17 41.24 48.21 56.89 54.34 53.33 More than one race, Non-Hispanic 51.91 70.72 69.56 80.47 71.97 73.45 Socioconomic status 27.64 47.23 64.45 64.40 69.56 47.53 67.32 Staq quintle 77.96 67.32 64.46 67.32 67.32 68.36 Staq quintle 77.96 83.61 87.19 76.73 87.18 76.18 Stad quintle 77.96 83.61 86.79 90.42 91.32 76.18 Stad quintle 77.96 88.31 86.79 90.42 91.32 76.18 Stad 77.96 88.31 86.79 90.42 91.32 76.32 School type 72.39 88.18 88.64 <td< td=""><td>Hispanic, race specified</td><td>40.27</td><td>38.63</td><td>61.51</td><td>62.22</td><td>68.20</td><td>67.14</td><td>58.44</td></td<>	Hispanic, race specified	40.27	38.63	61.51	62.22	68.20	67.14	58.44
Asian69.46 61.45 7.25 70.43 77.38 84.46 77.41 Hawaian, Oher Pacific Islander 53.36 39.18 56.40 54.06 70.70 82.50 69.70 American Indian, Alaska Native 34.17 34.17 41.24 48.21 56.89 54.34 53.38 More than one race, Non-Hispanic 57.78 51.91 70.72 69.56 80.47 71.97 73.45 Socioconomic status 28.62 27.04 42.41 49.97 54.66 47.53 42.23 Ist quintile (lowest) 28.62 27.04 42.41 49.97 54.66 47.53 42.23 Socioconomic status 28.62 27.04 42.41 49.97 54.66 47.53 42.23 Ist quintile (lowest) 28.62 57.99 55.95 79.17 77.99 78.08 68.35 Sid quintile 67.46 65.21 81.47 77.99 85.18 76.18 School type 57.99 55.95 79.17 77.99 85.18 76.16 School type 52.21 88.31 86.79 90.42 91.32 86.79 Public school 52.21 50.46 89.82 81.81 70.65 84.72 70.65 Public school 57.39 70.04 89.82 81.81 82.52 84.72 70.65	Hispanic, race not specified	27.67	30.11	46.36	55.68	66.41	64.64	53.79
Havaiian, Other Pacific Islander53.3639.1856.4054.0670.70 82.50 69.50American Indian, Alaska Native34.9734.1741.2448.2156.8954.3453.38American Indian, Alaska Native34.9734.1741.2448.2156.8954.3453.38More than one race, Non-Hispanic57.7851.9170.7269.56 80.47 71.9773.45Socioeconomic status28.6227.0442.4149.9754.6647.5342.23Lit quintile (novest)28.6227.0446.3364.4564.4069.5367.3254.86Socioeconomic status57.9955.9579.1772.7177.9978.0868.32Ath quintile67.4665.2181.4779.5983.6383.1876.18School type78.7076.6788.3186.7990.4291.3286.79Public school32.2150.4568.1887.8173.3971.6563.43Public school73.3970.0489.8281.8182.5284.7276.53Private school73.3970.0489.8281.8182.5284.7276.53Public school50.4568.6473.3981.8182.7284.7276.33Public school73.3970.0489.8281.8182.7284.7276.33Private school73.3970.0489.8281.8173.3971.6576.33<	Asian	69.46	61.45	72.25	70.43	77.38	84.46	77.41
American Indian, Alaska Native 34.97 34.17 41.24 48.21 56.89 54.34 53.38 More than one race, Non-Hispanic 57.78 51.91 70.72 69.56 80.47 71.97 73.45 Socioeconomic status 57.78 51.91 70.72 69.56 80.47 71.97 73.45 Socioeconomic status 28.62 27.04 42.41 49.97 54.66 47.53 42.23 Lat quintile (lowest) 28.62 27.04 45.33 64.45 64.40 69.53 67.32 54.86 Socioeconomic status 57.99 55.95 79.17 72.71 77.99 78.08 68.32 Ath quintile 67.46 65.21 81.47 79.59 83.63 85.18 76.18 School type 57.99 76.67 88.31 86.79 90.42 91.32 86.79 School type 52.21 50.45 68.18 86.64 73.39 71.65 63.43 Public school 73.39 70.04 89.82 81.81 82.52 84.72 76.67	Hawaiian, Other Pacific Islander	53.36	39.18	56.40	54.06	70.70	82.50	69.70
More than one race, Non-Hispanic 57.78 51.91 70.72 69.56 80.47 71.97 73.45 Socioeconomic status statuintile (lowest) 28.62 27.04 42.41 49.97 54.66 47.53 42.23 Socioeconomic status 14 quintile (lowest) 28.62 27.04 42.41 49.97 54.66 47.53 42.23 Socioeconomic status 47.09 46.33 64.45 64.40 69.53 67.32 54.86 Stat quintile (lowest) 57.99 55.95 79.17 72.71 77.99 78.08 68.32 Ath quintile (lowest) 78.70 76.67 88.31 86.79 90.42 91.32 86.70 School type 76.18 88.31 86.79 90.42 91.32 86.74 Public school 52.21 50.45 68.18 73.39 71.65 63.43 Public school 52.21 50.45 89.82 81.81 82.32 84.72 76.43 73.39 71.	American Indian, Alaska Native	34.97	34.17	41.24	48.21	56.89	54.34	53.38
Socioeconomic statusSocioeconomic status1st quintile (lowest) 28.62 27.04 42.41 49.97 54.66 47.53 42.23 2nd quintile (lowest) 28.62 27.04 42.41 49.97 54.66 47.53 42.23 2nd quintile (lowest) 57.99 55.95 79.17 72.71 77.99 78.08 68.32 4th quintile (highest) 78.70 76.67 88.31 81.47 79.59 83.63 85.18 76.18 5th quintile (highest) 78.70 76.67 88.31 86.79 90.42 91.32 86.70 School typePublic school 52.21 50.45 68.18 68.64 73.39 71.65 63.43 Public school 70.04 89.82 81.81 82.52 81.81 82.52 81.47	More than one race, Non-Hispanic	57.78	51.91	70.72	69.56	80.47	71.97	73.45
Ist quintile (lowest) 28.62 27.04 42.41 49.97 54.66 47.53 42.23 $2n d$ quintile 47.09 46.33 64.45 64.40 69.53 67.32 54.86 $3r d$ quintile 57.99 55.95 79.17 72.71 77.99 78.08 68.32 $4th$ quintile 67.46 65.21 81.47 79.59 83.63 85.18 76.18 $5th$ quintile (highest) 78.70 76.67 88.31 86.79 90.42 91.32 86.70 $5chool type$ 76.67 88.31 86.79 90.42 91.32 86.79 $8chool type$ 52.21 50.45 68.18 68.64 73.39 71.65 63.43 Public school 73.39 70.04 89.82 81.81 82.52 84.72 63.43	Socioeconomic status							
2nd quintile47.0946.33 64.45 64.40 69.53 67.32 54.86 3rd quintile57.9955.9579.1772.7177.9978.08 68.32 4th quintile 67.46 65.21 81.47 79.59 83.63 85.18 76.185th quintile (highest)78.7076.67 88.31 86.79 90.42 91.32 86.70 School typeSchool typeThis schoolThis school <td< td=""><td>1st quintile (lowest)</td><td>28.62</td><td>27.04</td><td>42.41</td><td>49.97</td><td>54.66</td><td>47.53</td><td>42.23</td></td<>	1st quintile (lowest)	28.62	27.04	42.41	49.97	54.66	47.53	42.23
3d quintile 57.99 55.95 79.17 72.71 77.99 78.08 68.32 $4th$ quintile 67.46 65.21 81.47 79.59 83.63 85.18 76.18 $5th$ quintile (highest) 78.70 76.67 88.31 86.79 90.42 91.32 86.70 $5th$ quintile (highest) 78.70 76.67 88.31 86.79 90.42 91.32 86.70 $5th$ quintile (highest) 78.70 76.67 88.31 86.79 90.42 91.32 86.70 $5th$ quintile (highest) 73.79 76.67 88.31 86.79 90.42 91.32 86.70 $5th$ quintile (highest) 73.39 70.04 88.31 86.79 90.42 91.32 86.70 8.51 87.2 88.18 88.64 73.39 71.65 63.43 76.73 70.04 89.82 81.81 82.52 84.72 76.73	2nd quintile	47.09	46.33	64.45	64.40	69.53	67.32	54.86
4th quintile 67.46 65.21 81.47 79.59 83.63 85.18 76.18 5 th quintile (highest) 78.70 76.67 88.31 86.79 90.42 91.32 86.70 6 chool type 76.04 88.31 86.79 90.42 91.32 86.70 86.00 type 72.21 50.45 68.18 68.64 73.39 71.65 63.43 Private school 73.39 70.04 89.82 81.81 82.52 84.72 76.73	3rd quintile	57.99	55.95	79.17	72.71	77.99	78.08	68.32
5th quintile (highest) 78.70 76.67 88.31 86.79 90.42 91.32 86.70 School type	4th quintile	67.46	65.21	81.47	79.59	83.63	85.18	76.18
School type School type 52.21 50.45 68.18 68.64 73.39 71.65 63.43 Public school 73.39 70.04 89.82 81.81 82.52 84.72 76.73	5th quintile (highest)	78.70	76.67	88.31	86.79	90.42	91.32	86.70
Public school 52.21 50.45 68.18 68.64 73.39 71.65 63.43 Private school 73.39 70.04 89.82 81.81 82.52 84.72 76.73	School type							
Private school 73.39 70.04 89.82 81.81 82.52 84.72 76.73	Public school	52.21	50.45	68.18	68.64	73.39	71.65	63.43
	Private school	73.39	70.04	89.82	81.81	82.52	84.72	76.73
	SOURCE: U.S. Department of Educati	ion, National Center fo	r Education Statistics,	Early Childhood Lon	gitudinal Study, Kind	ergarten Class of 1998	:-99 (ECLS-K), fall 19	998, spring 1999,
SUCRCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, SUCRCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999,	tall 1999, spring 2000, spring 2002, spi	ring, 2004, and spring 2	2007.					

Table A36. Percent of children at or above modal mathematics proficiency for each grade: School years 1998–99, 1999–2000, 2001–02,

	IF	AT paramete	srs	Used in		N	lean and star	ndard deviati	ion of theta ⁴		
Item label	a^1	b^2	c ³	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
CANDLE	0.79	-3.49	0.18	K1							
POURING	0.87	-2.76	0.17	K1							
CEREAL	1.15	-2.74	0.00	K1							
DECORATD	0.77	-2.63	0.15	K1							
BEGBIKE	1.65	-1.93	0.00	K1	*						
BEGIN	0.90	-1.85	0.00	K1	*						
LETRECD	2.71	-1.69	0.00	K1	*	*					
VEGETBLE	0.73	-1.66	0.13	K1	*	*					
LETRECF	3.08	-1.65	0.00	K1	*	*					
LETRECM	2.71	-1.64	0.00	K1	*	*					
LETRECT	2.89	-1.57	0.00	K1	*	*					
COULDNOT	0.90	-1.51	0.00	K1	*	*	*				
KAYLAFLY	0.66	-1.45	0.00	K1	*	*	*				
NEXTLINE	1.11	-1.42	0.00	K1	*	*	*				
STORYEND	1.28	-1.42	0.00	K1	*	*	*				
BEGP	1.74	-1.24	0.00	K1	-1.32	*	*				
TIME	1.07	-1.22	0.18	K1	(.51)	*	*				
BEGR	2.33	-1.21	0.00	K1	*	*	*				
BEGL	2.30	-1.18	0.00	K1	*	*	*				
TRUNK	0.82	-1.15	0.12	K1	*	*	*				
AWARDING	1.00	-1.07	0.28	K1	*	*	*				
JOGGING	1.23	-1.03	0.13	K1	*	*	*				
COULD	0.60	-1.03	0.00	K1	*	*	*				
ENDL	2.16	-1.00	0.00	K1	*	*	*				
MOM	2.36	-1.00	0.00	K1	*	*	*				
ENDF	1.80	-0.96	0.00	K 1	*	*	*				
See notes at end of table											

ECLS-K ITEM PARAMETERS BY ROUND

APPENDIX B

1 2006 07 1000 00 1000 Cohool . IDT :4 ÷ ρ Table R1

c' grades Round 1 (00) K1 K1 (10) K1 K1 (11) K1 K1 (12) K1 K1 (11) K1 K1 (12) K1 K1 (13) K1 K1 (14) K1 K1 (17) K1 K1 (18) 1 K1 (10) K1 K1 (11) K1 K1 (12) K1 K1 (13) K1 K1 (14) K1 K1 (15) K1 K1	Round 2 76 (.50) (.50) * * * * * * * * *	Round 3 	Kound 4 * * * * * * * * * * * * * * * * * * *	Kound 5 Ko	
000 000 000 000 000 000 000 000	* * * *	* * * * * * * * * * ;	* * * * * * * * * * * * *		
 (100 Ki (100 Ki (100	* * *	* * * * * * * \$ \$ \$ (.5)	* * * * * * * * * * * *		
 (100 (100 (100 (100 (100 (100 (110 (111 (112 (112 (113 (114 (114<td></td><td></td><td>* * * * * * * * * * * *</td><td></td><td></td>			* * * * * * * * * * * *		
 (00) (00)<td>76 () * * * * * * * * * * * * * * * * * * *</td><td></td><td>* * * * * * * * * * *</td><td></td><td></td>	76 () * * * * * * * * * * * * * * * * * * *		* * * * * * * * * * *		
 (00) (00)<td>76 (.50) (.58)</td><td></td><td>* * * * * * * * * *</td><td></td><td></td>	76 (.50) (.58)		* * * * * * * * * *		
 (00 (14 (00 (14 (14) <li< td=""><td>(0<u>5</u>.) (* * * * * * * * * * * *</td><td>* * * 5 (.51) * * * * * *</td><td>* * * * * * * * *</td><td></td><td></td></li<>	(0 <u>5</u> .) (* * * * * * * * * * * *	* * * 5 (.51) * * * * * *	* * * * * * * * *		
 (00 (00) (00) (00) (00) (10) (11) (12) (12) (13) (14) (15) (15) (16) (17) (19) (11) (11)<td>* * * * * * * * * *</td><td>52 * (.51) * * * * *</td><td>* * * * * * * *</td><td></td><td></td>	* * * * * * * * * *	52 * (.51) * * * * *	* * * * * * * *		
 (00 (00) (00) (00) (114) (00) (114) (11	* * * * * * * * * *	52 - .51) * * * * *	* * * * * * *		
 (00 (00 (00 (00 (14 (14 (14 (15 (16 (14 (14	* * * * * * * * *	52 (.51) * * * * * *	* * * * * *		
 (00 (100 (114 (114<td>* * * * * * * *</td><td>(.51)</td><td>* * * * *</td><td></td><td></td>	* * * * * * * *	(.51)	* * * * *		
 (00 K1,3 (00 K1,3 (14 K1 (14 K1 (10 K1,3 (17 K1 <li< td=""><td>* * * * * * *</td><td>* * * * *</td><td>* * * *</td><td></td><td></td></li<>	* * * * * * *	* * * * *	* * * *		
 (10) K1 (14) K1 (10) K1,3 (10) K1,3 (17) K1 (17) K1 (18) 17 (18) 117 (18) 117 (19) K1 (10) K1 (117) K1	* * * * * *	* * * *	* * * *		
 1.14 KI KI	* * * * *	* * *	* * •		
00 K1,3 19 K1 10 K1,3 10 K1,3 17 K1 17 K1 18 K1 117 K1 18 K1 10 K1	* * * *	* *	*		
00 KI,3 (19 KI,3 (00 KI,3 (17 KI (17 KI (18 I (18 I (18 I (18 I (19 KI (19 KI (* * *	*			
(19 K1 (00 K1,3 (22 K1 (17 K1 (17 K1 (18 1 (18 1 (11 1 (11 1))) (19 1) (19 1) (* *		*		
(00 KI,3 (00 KI (22 KI (17 KI (10 KI (18 I (18 I (10 KI (11 I (11 I (11)	*	*	*		
(00 K1 (22 K1 (17 K1 (100 K1 (18 1 (18 1) (18 1) (18 1) (19 1) (1		*	*		
	*	*	*		
0.17 K1 0.00 K1 0.18 1 0.00 K1 1	*	*	*		
.00 K1 1.18 1 0.00 K1	*	*	*		
18 1 000 K1	*	*	*		
0.00 K1	*	*	*		
	*	*	*		
000 K1	*	*	*		
000 K1	*	*	*		
0.50 K1	*	*	*		
0.25 K1	*	*	.08		
0.16 K1,3,5	*	*	(.47)		
0.15 K1	*	*	*	*	
0.25 K1	*	*	*	*	

	Round 7														*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
	Round 6									*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
on of theta ⁴	Round 5	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
idard deviatio	Round 4	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
ean and stan	Round 3	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*																
Μ	Round 2	*	*	*																												
	Round 1																															
Used in	grades	K1	K1,3,5	K1,3,5	K1	K1,3,5	1,3	K1	K1	ŝ	3,5	K1,3,5	ŝ	3,5	3,5	3,5	3,5	3,5	, m	K1,3	K1	1	3,5	1, 3, 5	3,5	K1,3	S	K1,3	3,5	1	1,3	
rs	c^3	0.00	0.12	0.19	0.13	0.13	0.00	0.16	0.00	0.12	0.29	0.00	0.14	0.21	0.31	0.33	0.00	0.11	0.09	0.00	0.19	0.00	0.18	0.00	0.00	0.00	0.11	0.00	0.25	0.00	0.00	
T parameter	b^2	0.17	0.17	0.23	0.26	0.27	0.36	0.37	0.38	0.42	0.43	0.43	0.45	0.47	0.50	0.50	0.51	0.51	0.51	0.52	0.53	0.53	0.53	0.53	0.55	0.56	0.57	0.58	0.59	0.60	0.60	
IR	a^1	3.65	3.96	3.75	4.00	4.56	3.57	2.80	3.74	2.69	3.33	2.84	2.15	2.53	1.97	3.78	3.43	1.59	1.80	3.55	4.00	2.65	4.38	5.82	2.88	2.53	2.90	2.65	3.83	2.64	4.00	
	Item label	WRONG	LISTEN	RIDEBIKE	CHOCCAKE	SIZES	QUIET	DOGHOUSE	ENVELOPE	COMPARSZ	KINDLETR	THROUGH	ADULT SZ	GROWUP	WHENPAST	WHENTOOK	GAVEWHAT	KNIGHT	DANGER	RAGE	MARCHED	CATNAME	AUTHFEEL	WTLESS	SAMEHANG	TOIL	KOALA IS	CORNER	TVSHOW	OWNRNAME	REQUIRE	

	Round 7	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
	Round 6	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
on of theta ⁴	Round 5	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	77.	(.32)	*	*	*	*	*	*	*	
dard deviatio	Round 4	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
ean and stan	Round 3																															
Μ	Round 2																															
	Round 1																															
Used in	grades	3,5	K1,3	3,5	K1,3	, ω	e	1, 3, 5	3,5	ω	1	S	3,5	, co	c	K1	3,5	5	5	3,5	3,5	∞	3,5	5	K1	5,8	3,5	3,5	3,5	ω	1	
s	c^3	0.00	0.00	0.00	0.00	0.24	0.00	0.00	0.15	0.13	0.00	0.00	0.19	0.00	0.16	0.19	0.00	0.17	0.16	0.00	0.00	0.16	0.00	0.16	0.19	0.27	0.17	0.00	0.15	0.00	0.00	
T parameter	b^2	0.61	0.62	0.64	0.64	0.65	0.66	0.66	0.67	0.67	0.68	0.69	0.69	0.71	0.71	0.71	0.72	0.73	0.74	0.74	0.76	0.76	0.77	0.78	0.78	0.79	0.79	0.79	0.80	0.85	0.86	
IR	a^1	3.62	2.97	2.99	2.06	3.68	4.09	4.03	3.43	2.14	4.00	1.55	2.81	2.02	2.87	3.90	2.14	1.79	3.86	2.82	2.23	1.69	3.33	3.45	5.14	1.87	3.50	3.09	4.42	2.31	3.95	
	Item label	TANZANIA	CAPTURE	KIND OFC	WEB	FACT	BABYSIT	MOISTURE	ROBBER	NONFICT	UNUSUAL	MOTHER	WORDARTH	STRAGGLE	TRUEBAB	RECIPE	THOSEDAY	MAINPROB	PREDICT	WHYROUND	JAMMED	PURPOSE	LINECLUE	EXAMPLE	INGREDNT	OVATIONS	BOW	SURPRISE	TRAIN	FRICTION	MYSTERLY	See notes at end of table

	Round 7	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
	Round 6	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
on of theta ⁴	Round 5	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
dard deviatio	Round 4	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
lean and stan	Round 3																															
A	Round 2																															
	Round 1																															
Used in	grades	1	5	5	∞	-	1,5	3,5	3,5	8	5,8	8	5,8	1	3,5	8	1	5,8	3,5	ŝ	5	ŝ	3,5,8	1	3,5	3,5	ω	5	5	3,5	3,5	
S	c^3	0.00	0.16	0.16	0.11	0.00	0.00	0.27	0.00	0.08	0.18	0.31	0.18	0.00	0.00	0.10	0.31	0.00	0.17	0.06	0.14	0.05	0.08	0.19	0.00	0.00	0.00	0.13	0.00	0.00	0.00	
T parameter	b^2	0.86	0.87	0.87	0.88	0.89	0.89	0.90	0.90	0.91	0.91	0.92	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.98	0.98	0.98	0.99	1.00	1.00	1.01	
IR	a ¹	1.95	3.49	2.17	3.71	2.61	3.55	3.34	3.02	1.70	3.18	3.16	3.97	4.05	2.55	1.46	2.34	2.64	2.01	3.55	3.07	2.84	1.62	2.72	2.93	2.98	2.12	3.18	3.07	2.75	1.85	
ſ	Item label	MOREINFO	DR ROSE	IMP UNDR	DECLINE	APPROX	WAGES	TEARING	FEELSAFE	NEWPLANS	DEPART	THEME	SPRING	VICIOUS	3THINGS	TRIALS	MNIDEA-S	4CORNERS	DEHYDRAT	POINT	SLUDGE	TAKECARE	HOWFEEL	ONYHW	DOMESTIC	LIKECHDR	APOSTRPH	DIFFRNT	SLOW LRN	INFLUENT	DIFFROOM	See notes at end of table

em label a' b' c' grades Round 1 Round 3 Round 3 Round 4 Round 5 Round 3 HYLEFT 3.82 1.01 0.00 1,35 4 <t< th=""><th></th><th>IR</th><th>tT paramete</th><th>STS</th><th>Used in</th><th></th><th>M</th><th>lean and star</th><th>ldard deviati</th><th>on of theta⁴</th><th></th><th></th></t<>		IR	tT paramete	STS	Used in		M	lean and star	ldard deviati	on of theta ⁴		
RITCISM 382 101 000 1.35 *	em label	\mathbf{a}^1	b^2	c^{3}	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
HYLEFT 3.63 1.01 0.00 5.8 ************************************	RITCISM	3.82	1.01	0.00	1,3,5				*	*	*	*
DMPASS 454 102 0.14 5 * 1.02 BOUT 3.12 102 0.14 5 * * 1.02 BFMONK 197 103 0.13 3.58 * * (.30) HYNOT 3.14 104 0.25 3.58 * (.30) HYNOT 3.14 104 0.25 3.58 * (.30) RANDS 1.83 0.01 1.3,5 * * (.30) RANDS 2.09 106 0.00 1.3,5 * (.31) RENUCE 2.09 106 0.01 3.5 * (.31) RENUCE 2.09 100 0.35 * * (.31) ATTHEME 2.36 1.11 0.11 3.5 * * (.31) ACHEL 2.71 1.12 0.20 3.5 * * (.31) ACHEL 2.77 1.11 0.11 <td>HY LEFT</td> <td>3.63</td> <td>1.01</td> <td>0.00</td> <td>5,8</td> <td></td> <td></td> <td></td> <td>*</td> <td>*</td> <td>*</td> <td>*</td>	HY LEFT	3.63	1.01	0.00	5,8				*	*	*	*
BOUT 3.12 1.02 0.12 5.5 ** (30) FFNONS 1.93 1.03 0.01 3.5.8 ** * (30) FFNONS 3.14 1.04 0.25 3.5.8 ** * (30) FFNNOS 3.14 1.04 0.25 3.5.8 ** * (30) SILKE 3.26 1.04 0.25 3.5.8 ** *<	OMPASS	4.54	1.02	0.14	5				*	*	1.02	*
M PROB 1.93 1.03 0.05 3.58 1.11 IFFMONK 1.97 1.03 0.01 3 3 1	BOUT	3.12	1.02	0.12	5				*	*	(.30)	*
FFMONK 1.97 1.03 0.11 3 HYNOT 3.14 1.04 0.25 3.5.8 HYNOT 3.14 1.04 0.25 3.5.8 RANDS 1.85 1.05 0.14 8 RENDE 3.16 1.05 0.01 8 RENDE 3.16 1.05 0.01 8 RENDE 2.09 1.06 0.00 1.3.5 REPROCE 2.09 1.06 0.00 3.5 ATTHEME 2.89 1.10 0.10 3.5 AUTHONT 2.67 1.10 0.10 3.5 AUTHONT 2.67 1.11 0.11 8 AUTHONT 2.67 1.11 0.11 8 AUTHONT 2.67 1.11 0.10 3.5 AUTHONT 2.67 1.11 0.00 3.5 AUTHONT 2.68 1.13 0.00 3.5 AURONT 2.73 1.12 0.00	M PROB	1.93	1.03	0.05	3,5,8					*	*	*
HYNOT 3:14 1.04 0.25 3,5.8 ************************************	IFFMONK	1.97	1.03	0.11	С					*	*	*
ISLIKE 3.26 1.04 0.12 8 ** ** RANDS 1.85 1.05 0.00 K1,3 ** ** ** ** JERVICED 3.16 1.05 0.01 8 ** ** ** ** ** JERVICE 2.09 1.06 0.01 3.5 **<	TONYH	3.14	1.04	0.25	3,5,8					*	*	*
TRANDS 1.85 1.05 0.00 K1.3 ************************************	ISLIKE	3.26	1.04	0.12	8					*	*	*
JRVIVED 3.16 1.05 0.14 8 ** ** REFINCE 209 1.06 0.00 1,3,5 ** *	FRANDS	1.85	1.05	0.00	K1,3					*	*	*
REFNCE 2.09 1.06 0.00 1,3,5 ************************************	JRVIVED	3.16	1.05	0.14	8					*	*	*
COBSOLV 191 1.07 0.00 3,5 ************************************	REFRNCE	2.09	1.06	0.00	1, 3, 5					*	*	*
ESCPURP 2.75 1.10 0.19 8 AJTHEME 2.89 1.10 0.10 3.5 ACHEL 2.71 1.12 0.01 3.5 ACHEL 2.71 1.12 0.00 3.5 HYCONTR 2.69 1.13 0.08 3.5 HYCONTR 2.69 1.13 0.08 3.5 MBITIOU 3.00 1.3 DPORTI 1.79 1.16 0.17 5.8 LIGNMNT 2.72 1.17 0.00 1.5 KAMS 2.73 1.17 0.00 1.5 KAMS 2.73 1.17 0.00 3.5 ELPPRB 2.78 1.17 0.00 3.5 ELPPRB 2.78 1.17 0.00 3.5 ELPPRB 2.78 1.17 0.00 3.5 COWAUTH 2.80 1.18 0.10 3.5 SI 1.17 0.00 3.5 COMPARMH 2.40 1.21 0.03 3.5 COMPARWH 2.40 1.21 0.03 3.5 COMPARWH 2.40 1.21 0.03 5.8 COMPARWH 2.40 1.21 0.01 3.5 COMPARWH 4.40 4.40 4.5 COMPARWH 4.40 4.40 4.5 COMPARWH	ROBSOLV	1.91	1.07	0.00	3,5					*	*	*
AJTHEME 2.89 1.10 0.10 3,5 AVNS 1.53 1.11 0.11 8 AVNS 1.53 1.11 0.11 8 AVNS 1.53 1.11 0.11 8 AVNS 1.53 1.11 0.11 8 AVNS 2.69 1.13 0.08 5 ICROWAV 2.97 1.14 0.00 3,5 MBITIOU 3.00 1.15 0.00 1,3 MBITIOU 3.00 1.15 0.00 1,3 MBITIOU 3.00 1.15 0.00 1,8 AVNS 2.73 1.17 0.07 3,5 LIGNMNT 2.73 1.17 0.07 3,5 AVNS 2.73 1.17 0.07 3,5 LIGNMNT 2.73 1.17 0.07 3,5 AVNS 2.73 1.17 0.00 1,5 AVNS 2.73 1.17 0.07 3,5 AVNS 2.73 1.17 0.00 1,5 AVNS 2.73 1.17 0.00 1,5 AVNANY 1.6 1.21 0.01 3,5 AVNS 2.32 1.21 0.01 3,5 AVNS 2.32 1.21 0.01 3,5 AVNANY 1.6 1.21 0.01 3,5 AVNA	ESCPURP	2.75	1.10	0.19	8					*	*	*
AWNS 1.53 1.11 0.11 8 ** ** ** ACHEL 2.71 1.12 0.20 3,5 ** ** ** ** HYCONTR 2.69 1.13 0.08 5 * ** ** ** HYCONTR 2.69 1.13 0.00 3,5 * ** ** ** ICROWAV 2.97 1.14 0.00 3,5 * ** ** ** JPPORTI 1.79 1.17 0.00 1,5 * ** ** ** LIGNMNT 2.72 1.17 0.00 1,5 * ** ** ** LIGNMNT 2.72 1.17 0.00 1,5 * ** ** ** VALOR 2.68 1.18 0.00 3,5 * ** ** ** ** OWAUTH 2.80 1.18 0.00 3,5 * ** ** ** ** OWAUTH 2.80 1.18 0.00 3,5 </td <td>AJTHEME</td> <td>2.89</td> <td>1.10</td> <td>0.10</td> <td>3,5</td> <td></td> <td></td> <td></td> <td></td> <td>*</td> <td>*</td> <td>*</td>	AJTHEME	2.89	1.10	0.10	3,5					*	*	*
ACHEL 2.71 1.12 0.20 3.5 **	AWNS	1.53	1.11	0.11	8					*	*	*
HYCONTR 2.69 1.13 0.08 5 **	ACHEL	2.71	1.12	0.20	3,5					*	*	*
ICROWAV 2.97 11.14 0.00 3,5 MBITIOU 3.00 11.5 0.00 1,3 JPPORT1 1.79 11.16 0.17 5,8 LIGNMNT 2.72 11.17 0.00 1,5 KAMS 2.73 11.17 0.00 8 ELPPRB 2.78 11.17 0.00 8 SUAUTH 2.80 11.18 0.00 3,5 N MESA 2.68 11.18 0.00 3,5 SUAUTH 2.80 11.18 0.00 3,5 N MESA 2.68 11.18 0.00 3,5 N MESA 2.68 11.18 0.00 3,5 N MATH 2.80 11.19 0.00 3,5 SUAUTH 2.80 11.19 0.00 3,5 N MATH 2.80 11.19 0.00 3,5 SUAUTH 2.80 11.19 0.00 3,5 N MATH 2.80 11.19 0.00 3,5 SUAUTH 2.80 11.19 0.00 3,5 N MATH 2.80 11.18 0.113 1 TPCOMP 3.20 11.19 0.00 3,5 SUAUTH 2.80 11.18 1 SUAUTH 2.80 11.18 0.113 1 SUAUTH 2.80 11.18 1 SUAUTH 2.80 1 SUAUTH 2.80 11.18 1 SUAUTH 2.80 1 SUAUTH	HYCONTR	2.69	1.13	0.08	5					*	*	*
MBITIOU 3.00 1.15 0.00 1,3 JPORTI 1.79 1.16 0.17 5,8 JPORTI 1.79 1.16 0.17 5,8 LIGNMNT 2.72 1.17 0.00 1,5 KAMS 2.73 1.17 0.00 1,5 KAMS 2.73 1.17 0.00 8 NEEA 2.68 1.18 0.00 5,8 NMESA 2.68 1.18 0.00 5,8 NAUTH 2.80 1.18 0.00 3,5 OWAUTH 2.80 1.18 0.00 3,5 SCRPIG 3.17 1.18 0.00 3,5 DWAUTH 2.80 1.18 0.00 3,5 OWAUTH 2.80 1.18 0.00 3,5 OWAUTH 2.80 1.18 0.00 3,5 OWAUTH 2.80 1.18 0.00 5 ELPOND 3.20 1.19 0.00 3,5 OMPARWH 2.40 1.21 0.00 3,5	ICROWAV	2.97	1.14	0.00	3,5					*	*	*
JPPORT1 1.79 1.16 0.17 5,8 ** ** ** LIGNMNT 2.72 1.17 0.00 1,5 ** ** ** ** ** KAMS 2.72 1.17 0.00 1,5 ** ** ** ** KAMS 2.73 1.17 0.00 8 ** ** ** ** WINSA 2.73 1.17 0.00 3,5,8 ** ** ** ** NMESA 2.68 1.18 0.00 3,5 ** ** ** ** OWAUTH 2.80 1.18 0.00 3,5 ** ** ** ** OWAUTH 2.80 1.18 0.00 3,5 ** ** ** ** OWAUTH 2.80 1.18 0.00 3,5 ** ** ** ** ** OWAUTH 2.80 1.18 0.00 3,5 ** ** ** ** ** EECRIG 3.17 1.18 0	MBITIOU	3.00	1.15	0.00	1,3					*	*	*
LIGNMNT 2.72 1.17 0.00 1,5 ** ** * KAMS 2.73 1.17 0.00 8 ** * * * KAMS 2.73 1.17 0.00 8 * * * * * ELPRB 2.73 1.17 0.00 8 * * * * * CLONTH 2.68 1.18 0.00 5,8 * * * * * * OWAUTH 2.80 1.18 0.00 3,5 *	JPPORT1	1.79	1.16	0.17	5,8					*	*	*
XAMS2.731.170.008**ELPRB2.781.170.073,5,8***ELPRB2.781.170.073,5,8***N MESA2.681.180.005,8****N MESA2.681.180.003,5*****OWAUTH2.801.180.003,5******SCRPIG3.171.180.005*******TPCOMP3.201.190.005*******TPCOMP3.201.190.005*******TPCOMP3.201.210.043,5*******OMPARWH2.401.210.035*******MMARY1.681.210.135*******	LIGNMNT	2.72	1.17	0.00	1,5					*	*	*
ELPPRB 2.78 1.17 0.07 3,5,8 ** </td <td>KAMS</td> <td>2.73</td> <td>1.17</td> <td>0.00</td> <td>∞</td> <td></td> <td></td> <td></td> <td></td> <td>*</td> <td>*</td> <td>*</td>	KAMS	2.73	1.17	0.00	∞					*	*	*
N MESA 2.68 1.18 0.00 5,8 * * * * * * * * * * * * * * * * * * *	ELPPRB	2.78	1.17	0.07	3,5,8					*	*	*
DWAUTH 2.80 1.18 0.00 3.5 *	N MESA	2.68	1.18	0.00	5,8					*	*	*
SCRPIG 3.17 1.18 0.13 1 *	DWAUTH	2.80	1.18	0.00	3,5					*	*	*
TPCOMP 3.20 1.19 0.00 5 *	ESCRPIG	3.17	1.18	0.13	1					*	*	*
JLPUND 2.32 1.21 0.04 3,5,8 *	TPCOMP	3.20	1.19	0.00	5					*	*	*
DMPARWH 2.40 1.21 0.00 3.5 *	ELPUND	2.32	1.21	0.04	3,5,8					*	*	*
JMMARY 1.68 1.21 0.13 5 * * *	DMPARWH	2.40	1.21	0.00	3,5					*	*	*
	JMMARY	1.68	1.21	0.13	5					*	*	*

	IR	T paramete	srs	Used in		M	fean and stan	idard deviati	on of theta ⁴		
Item label	a^1	b^2	c^3	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
LIKE/DIS	1.69	1.22	0.00	5					*	*	*
ERUPT2	2.75	1.23	0.00	S					*	*	*
AUTHOR	1.91	1.26	0.00	5					*	*	1.28
SHAREBOT	1.14	1.30	0.00	~					*	*	(.39)
PSYCHLG	1.81	1.31	0.00	5					*	*	*
TASKS	1.97	1.32	0.09	~					*	*	*
EFFORTS	2.30	1.32	0.10	~					*	*	*
PROTECT	2.13	1.33	0.24	8					*	*	*
HOAX	4.01	1.35	0.00	3,5					*	*	*
FEATURE	2.23	1.35	0.29	8					*	*	*
ADVANCES	1.44	1.36	0.00	5					*	*	*
DOUBT1	6.13	1.36	0.00	5,8					*	*	*
MNIDEA-A	4.16	1.36	0.18	3,5					*	*	*
INSUFFIC	2.47	1.38	0.00	5					*	*	*
GUESS	1.79	1.39	0.15	3,5					*	*	*
DOUBT2	5.64	1.40	0.00	5,8					*	*	*
TRUECROP	2.50	1.43	0.07	3,5						*	*
MAINPURP	3.24	1.48	0.28	5,8						*	*
THEORY2	1.37	1.48	0.00	5,8						*	*
?DISMISS	4.66	1.49	0.23	3,5						*	*
FIONA	1.64	1.50	0.26	~						*	*
BESTWAGM	3.28	1.57	0.21	3,5						*	*
TONE	3.02	1.59	0.17	5,8						*	*
RESULT	2.87	1.62	0.12	~						*	*
ALTRUIST	1.23	1.64	0.28	~							*
BELLGRNT	0.82	1.67	0.00	5							*
TRIFLES	2.00	1.68	0.10	8							*
ROBOTS	1.45	1.69	0.00	8							*
OVERLOOK	2.38	1.70	0.16	8							*
CATTLE	2.54	1.71	0.00	8							*
See notes at end of table											

	Π	RT paramete	rs	Used in		N	lean and star	idard deviati	on of theta ⁴		
Item label	a^1	b^2	c^3	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
ANOMALY	3.95	1.74	0.20	3							*
HISTORIAN	1.93	1.80	0.09	∞							*
SUPPORT2	2.04	1.92	0.13	8							*
CAUGHT	3.67	2.13	0.12	8							
EMBOLISM	1.16	2.43	0.00	ŝ							
RVLTIONS	1.09	2.52	0.15	8							
PROFLEV1	3.50	-1.57	0.00	K1							
PROFLEV2	3.26	-1.02	0.00	K1							
PROFLEV3	3.07	-0.73	0.00	K1							
PROFLEV4	4.00	-0.20	0.00	K1,3							
PROFLEV5	3.00	0.18	0.00	K1,3,5							
PROFLEV6	3.50	0.60	0.00	3,5							
PROFLEV7	4.00	0.86	0.00	3,5							
PROFLEV8	3.05	1.10	0.00	3,5,8							
PROFLEV9	5.92	1.55	0.00	5,8							
PROFLEV10	3.52	2.05	0.00	8							

¹ Parameter for discrimination.

² Parameter for difficulty.

³ Parameter for guessing.

⁴Mean and standard deviation (in parentheses) of theta ability estimate

overall difficulty (IRT "b" parameter). Four items not used in scale score are not included. The grades in which items appeared on assessment forms are noted. Mean and standard deviation of theta ability estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C4CW0, C5CW0, C5CW0, C5CW0, C7CW0). Asterisks mark the range corresponding to 2 standard NOTE: Item responses from kindergarten through eighth grade were pooled for IRT calibration to produce parameter estimates on a common scale. Items are sorted in estimated ascending order of deviations below and above the mean ability for the round.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

	IR	T paramet	ters	Used in			Mean and sta	indard devia	tion of theta ⁴		
Item label	a^1	b^2	c ³	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
2CRAYONS	1.91	-2.79	0.00	K1							
3BANANAS	0.96	-2.55	0.08	K1							
SQUARE	1.13	-2.39	0.14	K1							
NUMBER 4	3.50	-1.90	0.00	K1	*						
# STRAW	1.31	-1.79	0.00	K1	*						
STICKBAT	1.09	-1.68	0.05	K1	*						
3-1PENCL	0.99	-1.68	0.00	K1	*						
NUMBER 7	3.34	-1.66	0.00	K1	*						
#VANILLA	1.47	-1.60	0.00	K1	*	*					
#CHOC	1.55	-1.44	0.00	K1	*	*					
NUMBER 9	2.86	-1.43	0.00	K1	*	*					
PNTBRUSH	1.85	-1.36	0.21	K1	*	*	*				
COUNT 20	1.39	-1.32	0.00	K1	*	*	*				
4LINES	0.70	-1.29	0.17	K1	*	*	*				
6BANANAS	1.35	-1.23	0.00	K1	-1.19	*	*				
LG-SM-SM	1.80	-1.14	0.28	K1	(.48)	*	*				
SM-LG-SM	1.61	-1.13	0.23	K1	*	*	*				
NUMBER17	2.32	-1.05	0.00	K1	*	*	*				
000X	1.29	-0.99	0.19	K1	*	*	*				
NUMBER23	2.32	-0.91	0.00	K1	*	*	*				
3RD LINE	2.23	-0.88	0.00	K1	*	*	*				
3+2 CARS	1.56	-0.87	0.00	K1	*	*	*				
HALFOVAL	1.10	-0.85	0.22	K1	*	*	*				
78910	2.17	-0.85	0.00	K1	*	*	*				
$\overline{2}$ +3STICK	1.73	-0.81	0.00	K1	*	*	*	*			
#BUGS	1.69	-0.77	0.22	K1	*	*	*	*			
2 + 2	3.00	-0.75	0.00	K1	*	71	*	*			
3 + 3	4.00	-0.66	0.00	K1	*	(.46)	*	*			
1 + 7	1.61	-0.65	0.00	K1	*	*	*	*			
TEAMS_R	1.23	-0.64	0.15	3	*	*	*	*			
See notes at end of table.											

Table B2. Mathematics assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07
	IR	T paramet	ers	Used in			Mean and sta	ndard deviat	ion of theta ⁴		
Item label	a^{1}	b^2	c ³	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
VICKS R	2.62	-0.55	0.00	3	*	*	*	*			
8-6CRAYN	1.47	-0.53	0.00	K1	*	*	*	*			
3 + 4	2.47	-0.45	0.00	K1	*	*	46	*			
5-10RANG	2.15	-0.42	0.12	K1	*	*	(.48)	*			
2+5MARBL	1.55	-0.42	0.00	K1,3	*	*	*	*			
SHAPES	0.78	-0.36	0.20	K1	*	*	*	*			
PATTERN	1.58	-0.34	0.21	K1	*	*	*	*			
2+5CIRCL	1.83	-0.33	0.00	K1	*	*	*	*			
12 BY 2S	2.26	-0.31	0.00	K1,3	*	*	*	*			
3+7PENNY	2.25	-0.26	0.00	K1,3	*	*	*	*			
51015_25	2.53	-0.17	0.00	K1,3		*	*	*			
ORANGE R	1.90	-0.17	0.14	3		*	*	*			
11 + 3	2.49	-0.14	0.00	K1		*	*	*			
7 - 3	3.06	-0.12	0.00	K1		*	*	*			
9 - 2	3.21	-0.10	0.00	K1		*	*	*			
PATHS_R	1.26	-0.09	0.00	3		*	*	*	*		
6+7	2.52	-0.02	0.00	K1		*	*	.03	*		
12 + 6	2.07	0.07	0.00	K1		*	*	(.43)	*		
# MORE	2.14	0.09	0.00	K1		*	*	*	*		
MOST_Y	2.88	0.10	0.00	3		*	*	*	*		
2-1+2	1.96	0.12	0.00	K1		*	*	*	*		
RULERR	1.40	0.18	0.00	ю		*	*	*	*		
$A13_79$	1.95	0.18	0.00	K1,3,5		*	*	*	*		
SIDES_R	1.72	0.20	0.10	Э		*	*	*	*		
4+4-2	2.45	0.20	0.00	K1,3		*	*	*	*		
PAGES_R	2.53	0.21	0.20	3		*	*	*	*		
17 - 4	2.84	0.22	0.00	K1			*	*	*		
$COST_{10}$	2.46	0.23	0.00	K1,3,5			*	*	*	*	
12-9	2.79	0.24	0.00	K1			*	*	*	*	
26 + 20	2.90	0.29	0.00	K1			*	*	*	*	
See notes at end of table	ъ.										

Table B2. Mathematics assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07—Continued

7Continued
and 2006–07
2, 2003–04,
000, 2001-0
-99, 1999-2
l years 1998
eters: Schoo
l item param
cs assessment IR7
. Mathemati
Table B2.

	17												*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	Round																														
	Round 6	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
ion of theta ⁴	Round 5	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	.70	(.39)	*	*
ndard deviat	Round 4	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Mean and sta	Round 3	*	*	*	*	*	*	*	*	*	*	*																			
	Round 2																														
	Round 1																														
Used in	grades	K1,3,5	ŝ	ω	3,5	K1,3	K1,3,5	ŝ	K1,3	3,5	K1,3	K1	ŝ	ε	3,5	ŝ	3,5	ω	ω	K1,3	ω	ω	K1	ω	3,5	K1	3,5	3,5	ω	8	3,5
ers	c^3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
[paramete	b^2	0.30	0.33	0.34	0.36	0.40	0.41	0.44	0.45	0.45	0.46	0.47	0.51	0.52	0.52	0.52	0.53	0.55	0.56	0.57	0.58	0.60	0.63	0.64	0.65	0.66	0.68	0.70	0.70	0.71	0.71
IRJ	a ¹	2.58	2.72	1.04	1.01	1.73	2.53	4.38	2.77	2.74	1.33	3.01	3.06	3.23	2.44	3.72	2.19	2.01	3.38	2.47	3.04	2.63	3.08	3.90	3.66	2.39	1.45	3.12	3.91	2.93	2.15
	Item label	ARS15_5	FEWEST Y	SQUARE_R	CUBES10	HOWMANY\$	CANDY8 2	BEADS R	12-? PEN	NEXT78	HEADSUP	24-14BKS	MEANS R	EQUAL R	DO ADD4	MONEY R	TIME1030	POINTS R	SCORE Y	GOALS	PAPER	NICKELS	17CENTS	MORE1_Y	NUMBER60	BDCAKE	CUBESIDE	NEXT120	FEWERY	TREES -MC	CHART_64

ntinued
C
nd 2006–07-
04, a
2003-
-02, 2
2001-
-2000,
, 1999-
98–99.
ars 19
ol ye
: Scho
tem parameters:
RT i
s assessment I
Mathematic
Table B2.

	IR	T parame	ters	Used in			Mean and sta	ndard deviat	ion of theta ⁴		
Item label	a ¹	b^2	c^3	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
AGEGRAPH	1.82	0.72	0.00	5				*	*	*	*
BOX_700	3.67	0.73	0.00	3,5				*	*	*	*
NUMBER	2.69	0.74	0.00	3				*	*	*	*
TBSP7OZ	2.66	0.75	0.00	3,5,8				*	*	*	*
CANDY27	3.21	0.76	0.00	5				*	*	*	*
TREES100	3.02	0.77	0.00	5				*	*	*	*
COLORSYM	2.18	0.77	0.00	3,5				*	*	*	*
FRIES	3.01	0.77	0.00	3				*	*	*	*
CHILDR Y	2.82	0.78	0.00	3				*	*	*	*
STAR-Y	1.43	0.79	0.00	3				*	*	*	*
PGSLEFT	2.79	0.81	0.08	3,5,8				*	*	*	*
BOXSHELF	1.94	0.81	0.00	5				*	*	*	*
SECOND Y	2.65	0.82	0.00	3				*	*	*	*
SAMJUAN	4.01	0.82	0.11	8				*	*	*	*
A568214K	2.89	0.82	0.00	3,5				*	*	*	*
1ST#X5	2.19	0.82	0.22	5				*	*	*	*
BIKETIME	2.38	0.85	0.00	5				*	*	*	*
MISSNUM	1.61	0.87	0.00	5,8				*	*	*	*
FRUIT	2.05	0.87	0.12	3				*	*	*	*
24/4 TAB	1.74	0.89	0.00	K1				*	*	*	*
SCALE=	2.08	0.93	0.00	5					*	*	*
CARWASH	2.27	0.94	0.00	3,5,8					*	*	*
SHANBIKE	1.86	0.95	0.09	8					*	*	*
FLIPTRI	1.32	0.96	0.13	8					*	*	*
SCALEMAR	1.74	0.97	0.11	8					*	*	*
CARDS579	2.45	0.98	0.00	3,5					*	*	*
MARIA	2.29	0.98	0.00	3,5,8					*	*	*
24LEMONS	2.59	1.01	0.00	5					*	*	*
TILESCOV	1.62	1.02	0.00	ю					*	*	*
PROBNOT	1.94	1.02	0.15	8					*	*	*
See notes at end of table											

inued
Conti
Ĭ
,0- <u>9</u>
1200
, and
3–04
2003
-02,
001-
00, 2
)-20
1999
-99,
998
ars]
ol ye
Scho
ters:
ame
n par
iten
t IRJ
ment
ssess
ics a:
mati
1 athe
2
e B2
Tabl

	IR	T parame	sters	Used in		. 7	Mean and sta	ndard deviat	tion of theta ⁴		
Item label	a^1	b^2	c^3	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
100MORE	2.47	1.04	0.10	3,5,8					*	*	*
AREA_B	1.87	1.07	0.00	3					*	1.07	*
LARGER_B	2.01	1.10	0.00	3					*	.42	*
LOWDIVE	1.96	1.10	0.13	8					*	*	*
FOURLINE	1.53	1.12	0.18	8					*	*	*
TITO	2.09	1.13	0.00	3,5,8					*	*	*
PENCIL Y	1.29	1.14	0.05	3					*	*	*
LOUISA13	3.04	1.14	0.00	3,5					*	*	*
KYLEBOB	1.96	1.15	0.11	5,8					*	*	*
OPOS-MC	2.57	1.16	0.28	8					*	*	*
EQUAL B	2.13	1.16	0.10	3					*	*	*
AGE1/4	3.48	1.16	0.24	5					*	*	*
1/4STU44	4.16	1.17	0.00	5					*	*	*
LONGSTEP	2.05	1.24	0.00	5					*	*	*
7XPLUS4	2.12	1.24	0.19	8					*	*	*
NUMBE2_B	2.35	1.25	0.00	3					*	*	*
BEADSWHT	2.25	1.28	0.00	5					*	*	*
НОГГУ	2.02	1.29	0.00	3,5,8					*	*	*
TALL75_	2.95	1.29	0.10	3,5					*	*	*
X8ANDYX	1.91	1.30	0.25	8					*	*	*
YDOLLARS	1.81	1.31	0.08	8					*	*	*
CHANGE	2.00	1.31	0.00	K1					*	*	*
MARBLES	2.97	1.32	0.00	3,5					*	*	*
BANKER	2.00	1.33	0.00	3,5					*	*	*
SIDDORNG	2.65	1.36	0.22	8					*	*	*
MYSTER_B	2.79	1.37	0.00	3					*	*	*
NUMEDGES	1.45	1.38	0.00	3,5,8					*	*	*
OJ 300Z	2.72	1.39	0.00	5					*	*	*
FRAME3FT	2.70	1.40	0.00	5					*	*	1.41
SAND40MC	3.39	1.46	0.00	5,8					*	*	.46
See notes at end of table											

-Continued
-04, and 2006–07–
<u>99–2000, 2001–02, 2003</u>
ool years 1998–99, 199
ssment IRT item parameters: Sch
B2. Mathematics asse
[able]

	Round 7	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*						
	Round 6	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*										
ion of theta ⁴	Round 5																														
undard deviat	Round 4																														
Mean and sta	Round 3																														
	Round 2																														
	Round 1																														
Used in	grades	3,5	3,5,8	8	5,8	5	8	5,8	8	5,8	5	5	5,8	3	5	5	5,8	8	8	5	8	5,8	8	3,5,8	8	K1	K1	K1	K1,3	K1,3,5	
ers	c^3	0.00	0.00	0.13	0.21	0.11	0.15	0.11	0.00	0.04	0.00	0.21	0.15	0.00	0.00	0.09	0.00	0.12	0.16	0.00	0.14	0.00	0.05	0.00	0.12	0.00	0.00	0.00	0.00	0.00	
T paramet	b^2	1.50	1.50	1.51	1.53	1.53	1.56	1.57	1.57	1.57	1.59	1.60	1.64	1.70	1.71	1.71	1.73	1.74	1.83	1.84	1.89	1.94	2.03	2.25	2.31	-1.96	-1.24	-0.74	-0.18	0.40	
IR	a^1	2.32	2.08	2.71	3.16	3.05	2.80	2.85	1.89	2.66	3.00	1.42	2.79	2.41	4.00	2.81	2.73	2.65	2.28	2.51	1.51	3.20	1.98	2.03	1.71	3.50	3.29	4.00	3.93	4.77	0
	Item label	SAMEFRAC	MRKRULER	NUMSENT5	SMITHFAM	SHADED.2	X5PLUS3X	3FRAC4	LIFEEXP	RECPLAY	OPOSITIV	SALESTAX	PIZZA	FENCE B	SHADED.3	DIFF=88	VOLPRISM	500BATT	CROSSLIN	MEASDIAM	NPLUS5	CARPSFOE	CIRCCIRC	12INCH	FOLDCUBE	PROFLEV1	PROFLEV2	PROFLEV3	PROFLEV4	PROFLEV5	See notes at end of table

	IR	T paramet	ers	Used in			Mean and sta	andard deviat	tion of theta ⁴		
Item label	a^1	b^2	c ³	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
PROFLEV6	6.48	0.81	0.00	3,5							
PROFLEV7	4.68	1.20	0.00	3,5,8							
PROFLEV8	6.70	1.60	0.00	5,8							
PROFLEV9	6.33	1.93	0.00	5,8							
¹ Parameter for discrimination	ination										

Mathematics assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07—Continued Table B2.

rarameter for discrimination.

² Parameter for difficulty.

³ Parameter for guessing.

⁴ Mean and standard deviation (in parentheses) of theta ability estimate

NOTE: Item responses from kindergarten through eighth grade were pooled for IRT calibration to produce parameter estimates on a common scale. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). The grades in which items appeared on assessment forms are noted. Items whose format changed from multiple choice in grade 5 to open ended in grade 8 were not treated as common items, and are listed separately. Mean and standard deviation of theta ability estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0, C7CW0). Asterisks mark the range corresponding to 2 standard deviations below and above the mean ability for the round.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007.

	IRC	r paramete	ers	Used in		Δ	Aean and star	ndard deviati	on of theta ⁴		
Item label	a^1	\mathbf{b}^2	c ³	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
RBULB	0.91	-2.20	0.15	3							
RENRGY	1.30	-2.18	0.17	ю							
RPLANT	1.49	-2.12	0.17	Э							
RORGAN	0.58	-1.88	0.15	ω					*		
RTOOL	0.93	-1.83	0.12	ω					*		
ROUIMM	1.04	-1.82	0.03	3,5					*		
RDSAST	1.51	-1.75	0.09	ω					*		
RFGRPS	0.62	-1.65	0.23	3					*		
RFORMS	0.97	-1.63	0.17	3					*		
YPLAIN	0.79	-1.53	0.00	3					*		
RWINGS	1.63	-1.46	0.08	3,5					*		
RANIML	0.96	-1.39	0.10	3					*	*	
ROUFRZ	1.51	-1.30	0.10	3,5					*	*	
ROCCUR	1.59	-1.22	0.21	, κ					*	*	
WHCHPREY	1.26	-1.17	0.00	5					*	*	
RSEEDS	0.94	-1.12	0.08	3					*	*	
ROUTAP	0.62	-1.07	0.01	3,5					*	*	
ROUJUN	1.44	-1.07	0.00	3,5					*	*	
RTHING	1.21	-1.04	0.21	3					*	*	
RWATER	1.12	-1.03	0.10	ю					*	*	
YDSAST	0.83	-1.00	0.15	ω					*	*	
RSUNIS	1.23	-0.99	0.28	ω					*	*	
SOLAR	0.92	-0.94	0.15	8					*	*	
ROUBRN	1.44	-0.94	0.00	3,5					*	*	
ROUERT	0.88	-0.93	0.12	3,5,8					*	*	
RFISHB	1.06	-0.91	0.10	, S					*	*	
RSHAPE	1.14	-0.90	0.18	3					*	*	*
RHEART	1.35	-0.87	0.00	3,5					*	*	*
RPWDER	1.03	-0.86	0.15	ŝ					*	*	*
ROUJAR	0.63	-0.85	0.00	3,5					*	*	*
See notes at end of table.											

Table B3. Science assessment IRT item parameters: School years 2001–02, 2003–04, and 2006–07

Continued	
d 2006–07–	
2003–04, an	
rs 2001–02,	
School yea	
item parameters:	
Science assessment IRT i	
Table B3.	

	Round 7	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
	Round 6	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	08	(.68)	*	
on of theta ⁴	Round 5	*	*	*	*	66	(.67)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
ndard deviati	Round 4																															
Aean and star	Round 3																															
V	Round 2																															
	Round 1																															
Used in	grades	5,8	3,5	3,5	5,8	5	3,5,8	С	5	ω	ŝ	3,5,8	С	5,8	5,8	5	ŝ	3,5	5,8	5	5,8	3,5	3,5	3,5	3,5,8	3,5	3,5	∞	3,5,8	ω	5	
STS	c^3	0.12	0.42	0.18	0.05	0.00	0.09	0.00	0.00	0.10	0.19	0.29	0.00	0.23	0.16	0.15	0.16	0.17	0.18	0.00	0.15	0.00	0.21	0.07	0.23	0.25	0.13	0.14	0.08	0.12	0.00	
paramete	b^2	-0.78	-0.71	-0.69	-0.68	-0.66	-0.64	-0.62	-0.59	-0.53	-0.52	-0.45	-0.38	-0.38	-0.33	-0.32	-0.32	-0.32	-0.30	-0.27	-0.22	-0.19	-0.13	-0.13	-0.12	-0.12	-0.10	-0.10	-0.07	-0.06	-0.05	
IRT	a^1	1.06	1.24	1.05	0.68	1.08	0.73	0.77	0.99	0.99	1.27	1.38	1.04	1.48	1.17	1.30	1.13	0.85	1.28	1.51	0.88	1.13	0.90	1.14	1.60	1.51	1.35	0.66	0.93	0.80	0.96	
	Item label	CUTSCAB	ROUSRF	RDESRT	PYRAMID	MONSNTM	YTHEMT	BEARTH	SUGARDIS	VSOUND	YINSCT	YMOON	YSENSE	EARTHQK	GRAVMOON	PROTECT	BSHADW	ROUSOL	THUNDER	AIRPOLL	WATRGRPH	YBEES	ROUBLB	ROUGRT	ROUMTN	ROUMCE	ROUFLY	ECLIPSE8	BSOUND	YDSOLV	LAMPWIRE	See notes at end of table.

	IR'.	r paramete	ers	Used in		M	ean and star	idard deviati	on of theta ⁴		
Item label	a ¹	b^2	c^3	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
ROUSHD	0.87	0.03	0.00	3,5					*	*	*
MIXTURE	1.42	0.04	0.10	S					*	*	*
ECLIPSE5	1.15	0.04	0.00	5					*	*	*
YFWATE	1.41	0.05	0.24	ŝ					*	*	*
BPLNT2	1.04	0.09	0.11	3,5					*	*	*
YLIVE	1.50	0.17	0.12	ŝ					*	*	*
CUPTEMP	1.17	0.19	0.00	5					*	*	*
BPLANT	1.25	0.23	0.15	3,5					*	*	*
YBLANC	1.13	0.25	0.03	ŝ					*	*	*
BHIBER	0.76	0.28	0.22	ŝ					*	*	*
YFARMG	0.52	0.28	0.00	ŝ					*	*	*
EARTHCTR	0.90	0.32	0.25	8					*	*	*
BURIED	0.70	0.34	0.14	5,8					*	*	*
LIMESTON	1.02	0.36	0.25	8					*	*	*
BSLIDE	0.74	0.40	0.03	3,5,8					*	*	*
SEEDGROW	0.82	0.47	0.03	5,8					*	*	*
H2OSOURC	1.31	0.57	0.00	5					*	*	*
FOXRABIT	1.27	0.65	0.24	5,8					*	*	*
BSOIL	1.38	0.66	0.11	3,5					*	*	*
CHEMCHNG	0.66	0.67	0.15	5					*	*	*
BPLLUT	0.85	0.69	0.18	ŝ					*	*	*
BPOLAR	1.20	0.75	0.09	ŝ						*	*
BSTORM	1.81	0.76	0.19	ω						*	*
CONTOURX	0.75	0.85	0.16	8						*	.86
CAVITY	0.54	0.90	0.27	8						*	(.88)
YHUMID	0.75	0.91	0.10	ŝ						*	*
BPLNT3	1.35	0.93	0.06	ω						*	*
LAKE	1.04	1.00	0.20	8						*	*
TEMPMIX	1.25	1.04	0.10	8						*	*
PHYSPROP	1.41	1.06	0.24	5						*	*

Table B3. Science assessment IRT item parameters: School years 2001–02, 2003–04, and 2006–07—Continued

See notes at end of table.

Continued
d 2006–07–
, 2003–04, an
ears 2001–02
ers: School ye
item paramete
assessment IRT
Science
Table B3.

	Round 6 Round 7	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
on of theta ⁴	Round 5																					
Idard deviati	Round 4																					
lean and stan	Round 3																					
M	Round 2																					
	Round 1																					
Used in	grades	5	3,5	8	8	5	5	5,8	8	8	5	8	5,8	8	8	8	5	5	8	5	8	8
ers	c^3	00'0	0.00	0.29	0.28	0.18	0.25	0.14	0.24	0.17	0.00	0.35	0.13	0.14	0.23	0.14	0.24	0.00	0.09	0.31	0.17	0.06
paramete	b^2	1.12	1.13	1.14	1.26	1.32	1.36	1.41	1.41	1.42	1.42	1.43	1.44	1.53	1.61	1.63	1.79	2.02	2.07	2.32	2.32	2.47
IRI	a^1	1.03	0.78	1.05	1.16	0.99	1.16	0.84	1.23	1.21	0.78	0.67	0.88	0.98	0.94	0.91	0.52	0.57	1.24	0.43	0.84	0.68
	Item label	NERVOUS	BMAMML	PROPERTY	CREEK	SUNMOVE	CONSTELL	SOLUTION	PREDATOR	ALGAE	TEMPLOW	CELLS	PENCLH20	REPRODUC	PRECIPIT	STORM	BEARCUB	WHYFAST	GRANITE	H2ORECYC	LENS	GUINEA

¹ Parameter for discrimination.

² Parameter for difficulty. ³ Parameter for guessing.

⁴ Mean and standard deviation (in parentheses) of theta ability estimate

NOTE: Item responses from third through eighth grade were pooled for IRT calibration to produce parameter estimates on a common scale. Science was not tested in kindergarten/first grade. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). The grades in which items appeared on assessment forms are noted. Mean and standard deviation of theta ability estimates are based on cross-sectional weights within each round (C5CW0, C6CW0, C7CW0). Asterisks mark the range corresponding to 2 standard deviations below and above the mean ability for the round. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2002, spring 2004, and spring 2007.

2006–07	
14, and 2	l
)3–0	
02, 20(٩
, 2001–	
9-2000	ŗ
199	,
-99,	•
1998	¢
years	, ,
lood	
t: Sc	ŕ
rrect	
n co	
portio	¢
l prc	
lated	
stin	(
ent e	
ssme	;
asse	
ling	
Reac	
C1.	
Table	

ECLS-K ESTIMATED PROPORTION CORRECT BY ROUND

APPENDIX C

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
CANDLE	K1	0.95	0.98	0.98	0.99	1.00	1.00	1.00
POURING	K1	0.90	0.95	0.96	0.98	0.99	1.00	1.00
CEREAL	K1	0.92	0.97	0.98	0.99	1.00	1.00	1.00
DECORATD	K1	0.85	0.92	0.94	0.97	0.99	0.99	0.99
BEGBIKE	K1	0.78	0.93	0.96	0.99	1.00	1.00	1.00
BEGIN	K1	0.67	0.82	0.86	0.94	0.98	0.99	0.99
LETRECD	K1	0.71	0.93	0.96	0.99	1.00	1.00	1.00
VEGETBLE	K1	0.65	0.77	0.81	0.90	0.96	0.97	0.98
LETRECF	K1	0.69	0.93	0.96	0.99	1.00	1.00	1.00
LETRECM	K1	0.68	0.92	0.96	0.99	1.00	1.00	1.00
LETRECT	K1	0.64	0.90	0.95	0.99	1.00	1.00	1.00
COULDNOT	K1	0.56	0.74	0.79	0.90	0.97	0.98	0.98
KAYLAFLY	K1	0.53	0.68	0.73	0.84	0.92	0.94	0.95
NEXTLINE	K1	0.53	0.74	0.81	0.92	0.98	0.99	0.99
STORYEND	K1	0.54	0.76	0.83	0.94	0.99	0.99	1.00
BEGP	K1	0.45	0.74	0.83	0.95	1.00	1.00	1.00
TIME	K1	0.56	0.73	0.79	0.91	0.97	0.98	0.99
BEGR	K1	0.42	0.76	0.85	0.97	1.00	1.00	1.00
BEGL	K1	0.40	0.74	0.84	0.96	1.00	1.00	1.00
TRUNK	K1	0.51	0.67	0.73	0.85	0.94	0.96	0.97
AWARDING	K1	0.57	0.72	0.78	0.89	0.97	0.98	0.98
JOGGING	K1	0.45	0.67	0.74	0.89	0.98	0.99	0.99
COULD	K1	0.43	0.57	0.62	0.75	0.86	0.89	0.91
ENDL	K1	0.31	0.64	0.76	0.94	1.00	1.00	1.00
MOM	K1	0.30	0.65	0.77	0.94	1.00	1.00	1.00
See notes at end of table.								

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
ENDF	K1	0.31	0.61	0.72	0.92	0.99	1.00	1.00
YELLOW	K1	0.27	0.58	0.70	0.91	0.99	1.00	1.00
BEGB	K1	0.30	0.56	0.66	0.87	0.98	0.99	0.99
BEGWORD	K1	0.35	0.53	0.61	0.78	0.91	0.93	0.95
ENDP	K1	0.26	0.53	0.65	0.87	0.98	0.99	0.99
QMARK	K1	0.30	0.52	0.61	0.81	0.94	0.96	0.98
ENDD	K1	0.21	0.46	0.59	0.84	0.98	0.99	0.99
YOU	K1	0.15	0.45	0.60	0.89	1.00	1.00	1.00
ORPIG	K1	0.14	0.38	0.52	0.83	0.98	0.99	1.00
ORSAT	K1	0.11	0.35	0.50	0.85	0.99	1.00	1.00
ORTAIL	K1	0.08	0.29	0.44	0.82	0.99	1.00	1.00
RUNS	K1,3	0.06	0.26	0.40	0.81	0.99	1.00	1.00
ORHAND	K1	0.06	0.25	0.39	0.80	0.99	1.00	1.00
NEEDHOME	K1	0.18	0.31	0.43	0.81	0.99	1.00	1.00
WENT	K1,3	0.05	0.20	0.33	0.76	0.98	1.00	1.00
DOWN	K1,3	0.04	0.18	0.30	0.74	0.98	1.00	1.00
BOYBIKE	K1	0.22	0.33	0.43	0.79	0.99	1.00	1.00
JEEP	K1,3	0.04	0.18	0.30	0.72	0.98	0.99	1.00
FISHING	K1	0.03	0.14	0.26	0.72	0.98	1.00	1.00
CANINBAG	K1	0.27	0.38	0.46	0.74	0.96	0.98	0.99
KITNBED	K1	0.20	0.30	0.39	0.75	0.98	0.99	1.00
CATCH	K1	0.03	0.14	0.25	0.70	0.97	0.99	1.00
MAKE	1	0.26	0.36	0.43	0.66	0.88	0.93	0.95
KNOW	K1	0.03	0.12	0.20	0.59	0.93	0.98	0.99
LIGHT	K1	0.02	0.08	0.16	0.59	0.96	0.99	1.00
ELEPHANT	K1	0.02	0.08	0.16	0.58	0.95	0.99	1.00
KIMCAT	K1	0.51	0.54	0.58	0.78	0.98	0.99	1.00
GIRLRED	K1	0.26	0.31	0.37	0.65	0.95	0.98	0.99
BACKPACK	K1,3,5	0.17	0.22	0.28	0.59	0.93	0.98	0.99
FLATTIRE	K1	0.16	0.20	0.26	0.56	0.93	0.98	0.99
LIKEDRY	K1	0.26	0.29	0.34	0.61	0.95	0.99	1.00
WRONG	K1	0.01	0.05	0.11	0.46	0.92	0.97	0.99
See notes at end of table.								

Continued
nd 2006–07—
001-02, 2003-04, a
9, 1999–2000, 20
years 1998–9
sct: School
ated proportion corre
ading assessment estima
Table C1. Rei

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
K1,3,5		0.13	0.17	0.22	0.53	0.93	0.98	0.99
K1,3,5		0.20	0.23	0.27	0.53	0.92	0.97	0.99
K1		0.14	0.17	0.21	0.47	0.91	0.97	0.99
K1,3,5		0.14	0.16	0.20	0.46	0.91	0.97	0.99
1,3		0.01	0.03	0.07	0.32	0.84	0.94	0.97
Kl		0.17	0.20	0.23	0.43	0.83	0.93	0.96
Kl		0.01	0.03	0.07	0.30	0.83	0.94	0.97
ŝ		0.13	0.15	0.19	0.38	0.80	0.90	0.95
3,5		0.29	0.30	0.33	0.48	0.85	0.94	0.97
K1,3,5		0.01	0.03	0.07	0.28	0.77	0.89	0.94
ю		0.15	0.18	0.22	0.40	0.76	0.87	0.92
3,5		0.22	0.23	0.26	0.42	0.79	0.89	0.94
3,5		0.32	0.34	0.37	0.50	0.78	0.87	0.92
3,5		0.34	0.35	0.36	0.48	0.84	0.93	0.97
3,5		0.00	0.02	0.05	0.22	0.74	0.89	0.94
3,5 (J	0.13	0.17	0.20	0.37	0.69	0.80	0.87
3 (U	0.10	0.14	0.17	0.35	0.70	0.81	0.88
K1,3 (Ŭ	00.0	0.02	0.05	0.22	0.74	0.88	0.94
K1 (U	0.19	0.20	0.23	0.36	0.79	0.91	0.96
1	-	0.01	0.03	0.06	0.23	0.70	0.84	0.91
3,5		0.18	0.19	0.21	0.34	0.79	0.92	0.96
1,3,5		0.00	0.01	0.04	0.18	0.76	0.92	0.96
3,5		0.01	0.02	0.05	0.21	0.69	0.84	0.91
K1,3		0.01	0.02	0.05	0.22	0.67	0.82	06.0
ю		0.11	0.12	0.15	0.29	0.71	0.85	0.92
K1,3		0.01	0.02	0.05	0.20	0.66	0.82	0.90
3,5		0.25	0.26	0.28	0.38	0.76	0.89	0.94
1		0.01	0.02	0.05	0.20	0.65	0.81	0.89
1,3		0.00	0.01	0.03	0.16	0.68	0.85	0.92
3,5		0.00	0.01	0.03	0.16	0.67	0.84	0.92
K1,3		0.00	0.02	0.04	0.17	0.64	0.81	0.89
3,5		0.00	0.01	0.04	0.16	0.62	0.80	0.89

-Continued
and 2006–07—
01-02, 2003-04,
9, 1999–2000, 20
ears 1998–99
ect: School y
estimated proportion corr
Reading assessment
Table C1.

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
WEB	K1,3	0.01	0.03	0.06	0.20	09.0	0.75	0.84
FACT	, ε	0.24	0.25	0.26	0.35	0.72	0.86	0.93
BABYSIT	3	0.00	0.01	0.03	0.13	0.63	0.82	0.90
MOISTURE	1, 3, 5	0.00	0.01	0.03	0.13	0.62	0.82	0.90
ROBBER	3,5	0.15	0.16	0.18	0.27	0.67	0.83	0.91
NONFICT	3	0.14	0.15	0.18	0.30	0.64	0.77	0.86
UNUSUAL	1	0.00	0.01	0.03	0.12	0.61	0.81	0.90
MOTHER	5	0.01	0.04	0.08	0.22	0.55	0.68	0.79
WORDARTH	3,5	0.20	0.20	0.22	0.31	0.66	0.81	0.89
STRAGGLE	3	0.01	0.02	0.05	0.18	0.55	0.71	0.82
TRUEBAB	3	0.16	0.17	0.18	0.27	0.63	0.79	0.88
RECIPE	K1	0.19	0.20	0.21	0.28	0.66	0.83	0.90
THOSEDAY	3,5	0.01	0.02	0.04	0.17	0.54	0.71	0.82
MAINPROB	5	0.18	0.20	0.22	0.33	0.61	0.74	0.83
PREDICT	5	0.16	0.17	0.18	0.24	0.62	0.80	0.89
WHYROUND	3,5	0.00	0.01	0.03	0.13	0.54	0.73	0.84
JAMMED	3,5	0.00	0.02	0.04	0.15	0.52	0.69	0.81
PURPOSE	8	0.17	0.19	0.21	0.31	0.59	0.71	0.81
LINECLUE	3,5	0.00	0.01	0.02	0.10	0.51	0.72	0.84
EXAMPLE	5	0.17	0.17	0.18	0.24	0.59	0.77	0.87
INGREDNT	K1	0.19	0.19	0.20	0.24	0.60	0.80	0.89
OVATIONS	5,8	0.27	0.28	0.30	0.38	0.63	0.74	0.83
BOW	3,5	0.17	0.17	0.18	0.24	0.58	0.76	0.86
SURPRISE	3,5	0.00	0.01	0.02	0.10	0.49	0.70	0.82
TRAIN	3,5	0.15	0.15	0.16	0.21	0.56	0.76	0.87
FRICTION	ŝ	0.00	0.01	0.03	0.11	0.45	0.63	0.76
MYSTERLY	1	0.00	0.00	0.01	0.06	0.43	0.66	0.81
MOREINFO	1	0.00	0.02	0.03	0.13	0.44	0.61	0.74
DR ROSE	5	0.16	0.16	0.17	0.21	0.51	0.70	0.83
IMP UNDR	5	0.16	0.17	0.18	0.25	0.52	0.67	0.79
DECLINE	8	0.11	0.12	0.12	0.17	0.48	0.68	0.82
APPROX	1	0.00	0.01	0.02	0.09	0.42	0.61	0.76
See notes at end of table.								

Continued
and 2006–07–
)2, 2003–04, ;
-2000, 2001–(
998–99, 1999-
School years 1
rtion correct:
estimated propo
Reading assessmen
Table C1.

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
WAGES	1,5	0.00	0.00	0.01	0.06	0.40	0.62	0.78
TEARING	3,5	0.27	0.27	0.28	0.32	0.56	0.72	0.84
FEELSAFE	3,5	0.00	0.01	0.01	0.07	0.40	0.61	0.76
NEWPLANS	8	0.08	0.10	0.11	0.20	0.46	0.60	0.72
DEPART	5,8	0.18	0.18	0.19	0.23	0.49	0.67	0.80
THEME	8	0.31	0.31	0.32	0.35	0.57	0.72	0.83
SPRING	5,8	0.18	0.18	0.19	0.21	0.46	0.65	0.80
VICIOUS	1	0.00	0.00	0.01	0.04	0.34	0.57	0.76
3THINGS	3,5	0.00	0.01	0.02	0.07	0.37	0.56	0.72
TRIALS	8	0.10	0.12	0.14	0.23	0.46	0.58	0.69
MNIDEA-S	1	0.31	0.32	0.32	0.37	0.56	0.69	0.80
4CORNERS	5,8	0.00	0.01	0.01	0.07	0.36	0.55	0.72
DEHYDRAT	3,5	0.18	0.18	0.19	0.25	0.49	0.62	0.75
POINT	3	0.06	0.07	0.07	0.10	0.38	0.59	0.76
SLUDGE	5	0.14	0.15	0.15	0.19	0.44	0.62	0.77
TAKECARE	3	0.05	0.05	0.06	0.10	0.37	0.57	0.73
HOWFEEL	3,5,8	0.09	0.10	0.12	0.20	0.44	0.57	0.69
ONYHW	1	0.19	0.19	0.20	0.23	0.47	0.63	0.77
DOMESTIC	3,5	0.00	0.00	0.01	0.05	0.33	0.54	0.72
LIKECHDR	3,5	0.00	0.00	0.01	0.05	0.33	0.54	0.71
APOSTRPH	ς	0.00	0.01	0.02	0.08	0.36	0.53	0.68
DIFFRNT	5	0.13	0.13	0.13	0.17	0.40	0.59	0.75
SLOW LRN	5	0.00	0.00	0.01	0.04	0.31	0.52	0.70
INFLUENT	3,5	0.00	0.00	0.01	0.05	0.32	0.51	0.69
DIFFROOM	3,5	0.00	0.01	0.03	0.10	0.35	0.51	0.66
CRITCISM	1,3,5	0.00	0.00	0.01	0.03	0.28	0.51	0.71
WHY LEFT	5,8	0.00	0.00	0.01	0.03	0.29	0.51	0.70
COMPASS	S	0.14	0.14	0.14	0.16	0.37	0.57	0.75
ABOUT	5	0.12	0.12	0.12	0.15	0.37	0.55	0.72
SIM PROB	3,5,8	0.05	0.06	0.07	0.13	0.37	0.52	0.67
DIFFMONK	3	0.11	0.12	0.13	0.18	0.41	0.55	0.69
WHYNOT	3,5,8	0.25	0.26	0.26	0.28	0.46	0.62	0.76
See notes at end of table.								

-Continued	
and 2006–07–	
1-02, 2003-04,	
1999–2000, 200	
ears 1998–99,	
rrect: School y	
imated proportion co	
Reading assessment est	
Table C1.	

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
DISLIKE	8	0.12	0.12	0.12	0.15	0.36	0.55	0.72
STRANDS	K1,3	0.00	0.01	0.02	0.09	0.33	0.48	0.63
SURVIVED	8	0.14	0.14	0.15	0.17	0.37	0.55	0.72
PREFRNCE	1,3,5	0.00	0.01	0.02	0.07	0.31	0.47	0.64
PROBSOLV	3,5	0.00	0.01	0.02	0.08	0.31	0.47	0.63
DESCPURP	8	0.19	0.19	0.20	0.22	0.39	0.54	0.70
MAJTHEME	3,5	0.10	0.10	0.11	0.13	0.32	0.48	0.67
PAWNS	8	0.12	0.13	0.14	0.20	0.39	0.51	0.64
RACHEL	3,5	0.20	0.20	0.21	0.23	0.39	0.53	0.69
WHYCONTR	5	0.08	0.08	0.09	0.11	0.29	0.45	0.64
MICROWAV	3,5	0.00	0.00	0.01	0.03	0.21	0.39	0.60
AMBITIOU	1,3	0.00	0.00	0.01	0.02	0.20	0.38	0.60
SUPPORT1	5,8	0.17	0.18	0.19	0.23	0.39	0.51	0.64
ALIGNMNT	1,5	0.00	0.00	0.01	0.03	0.21	0.38	0.58
EXAMS	8	0.00	0.00	0.01	0.03	0.20	0.37	0.58
HELPPRB	3,5,8	0.07	0.07	0.07	0.09	0.26	0.41	0.61
ON MESA	5,8	0.00	0.00	0.01	0.03	0.20	0.37	0.57
HOWAUTH	3,5	0.00	0.00	0.01	0.03	0.19	0.36	0.57
DESCRPIG	1	0.13	0.13	0.13	0.14	0.28	0.43	0.63
MTPCOMP	5	0.00	0.00	0.00	0.02	0.17	0.35	0.57
HELPUND	3,5,8	0.05	0.05	0.05	0.08	0.24	0.39	0.57
COMPARWH	3,5	0.00	0.00	0.01	0.03	0.20	0.35	0.55
SUMMARY	5	0.13	0.14	0.14	0.19	0.35	0.46	0.60
LIKE/DIS	5	0.00	0.01	0.02	0.06	0.25	0.38	0.53
ERUPT2	5	0.00	0.00	0.00	0.02	0.17	0.33	0.54
AUTHOR	5	0.00	0.01	0.01	0.05	0.21	0.34	0.51
SHAREBOT	8	0.01	0.03	0.04	0.11	0.28	0.38	0.49
PSYCHLG	5	0.00	0.01	0.01	0.05	0.20	0.32	0.48
TASKS	8	0.09	0.09	0.09	0.12	0.25	0.36	0.52
EFFORTS	8	0.10	0.10	0.10	0.12	0.23	0.35	0.52
PROTECT	8	0.24	0.24	0.25	0.26	0.36	0.46	0.60
HOAX	3,5	0.00	0.00	0.00	0.00	0.07	0.20	0.44
See notes at end of table.								

)			3	<u>,</u>				
	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
FEATURE	8	0.29	0.29	0.29	0.30	0.39	0.48	0.61
ADVANCES	5	0.00	0.01	0.02	0.07	0.22	0.32	0.46
DOUBT1	5,8	0.00	0.00	0.00	0.00	0.04	0.16	0.43
MNIDEA-A	3,5	0.18	0.18	0.18	0.19	0.23	0.34	0.53
INSUFFIC	5	0.00	0.00	0.00	0.02	0.12	0.24	0.43
GUESS	3,5	0.15	0.16	0.16	0.18	0.29	0.39	0.52
DOUBT2	5,8	0.00	0.00	0.00	0.00	0.03	0.14	0.39
TRUECROP	3,5	0.07	0.07	0.07	0.08	0.16	0.26	0.44
MAINPURP	5,8	0.28	0.28	0.29	0.29	0.32	0.39	0.54
THEORY2	5,8	0.00	0.01	0.02	0.06	0.18	0.27	0.40
3DISMISS	3,5	0.23	0.23	0.23	0.23	0.25	0.31	0.48
FIONA	8	0.26	0.26	0.27	0.28	0.36	0.43	0.54
BESTWAGM	3,5	0.21	0.21	0.21	0.21	0.24	0.29	0.43
TONE	5,8	0.17	0.17	0.17	0.17	0.20	0.26	0.40
RESULT	8	0.12	0.12	0.12	0.13	0.16	0.21	0.35
ALTRUIST	8	0.29	0.29	0.30	0.32	0.39	0.45	0.53
BELLGRNT	5	0.02	0.04	0.06	0.11	0.23	0.30	0.38
TRIFLES	8	0.11	0.11	0.11	0.11	0.16	0.22	0.34
ROBOTS	8	0.00	0.01	0.01	0.03	0.11	0.18	0.30
OVERLOOK	8	0.16	0.16	0.16	0.16	0.19	0.24	0.36
CATTLE	8	0.00	0.00	0.00	0.00	0.03	0.08	0.22
ANOMALY	3	0.20	0.20	0.20	0.20	0.20	0.22	0.33
HISTORIAN	8	0.09	0.09	0.09	0.10	0.13	0.18	0.28
SUPPORT2	8	0.13	0.13	0.13	0.14	0.16	0.19	0.27
CAUGHT	8	0.12	0.12	0.12	0.12	0.12	0.13	0.16
EMBOLISM	ς	0.00	0.00	0.00	0.01	0.04	0.07	0.11
RVLTIONS	8	0.15	0.15	0.15	0.16	0.18	0.20	0.24
NOTE: IRT-estimated proportio estimated ascending order of ov items appeared in test forms for a	n correct for each item in eac erall difficulty (IRT "b" paran all rounds. Table estimates are	h round. Estimates f acter). Four items no based on cross secti	Or kindergarten thro ot used in scale scor onal-weights within	ugh eighth grade ha e are not included. T each round (C1CW0	ve been put on a cor The grades in which), C2CW0, C3CW0,	nmon scale to suppo items appeared on a C4CW0, C5CW0, C6	rt comparisons. Iter assessment forms are 5CW0, C7CW0).	ns are sorted in noted. Not all
SOURCE: U.S. Department of I spring 2000, spring 2002, spring	Education, National Center for 2004, and spring 2007.	r Education Statistic	ss, Early Childhood	Longitudinal Study,	Kindergarten Class	of 1998-99 (ECLS-K	.), fall 1998, spring	1999, fall 1999,

Table C1. Reading assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07—Continued

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
2CRAYONS	K1	0.98	1.00	1.00	1.00	1.00	1.00	1.00
3BANANAS	K1	0.89	0.94	0.96	0.98	0.99	1.00	1.00
SQUARE	K1	0.90	0.95	0.97	0.99	1.00	1.00	1.00
NUMBER 4	K1	0.90	0.98	0.99	1.00	1.00	1.00	1.00
# STRAW	K1	0.75	0.88	0.92	0.97	0.99	1.00	1.00
STICKBAT	K1	0.70	0.84	0.89	0.95	0.99	0.99	1.00
3-1PENCL	K1	0.67	0.81	0.86	0.93	0.98	0.99	0.99
NUMBER 7	K1	0.80	0.95	0.98	1.00	1.00	1.00	1.00
#VANILLA	K1	0.69	0.86	0.91	0.97	0.99	1.00	1.00
#CHOC	K1	0.62	0.82	0.88	0.96	0.99	1.00	1.00
NUMBER 9	K1	0.66	0.89	0.94	0.99	1.00	1.00	1.00
PNTBRUSH	K1	0.68	0.85	0.91	0.97	1.00	1.00	1.00
COUNT 20	K1	0.56	0.76	0.84	0.94	0.99	0.99	1.00
4LINES	K1	0.61	0.71	0.76	0.85	0.92	0.95	0.96
6BANANAS	K1	0.52	0.73	0.81	0.92	0.98	0.99	1.00
LG-SM-SM	K1	0.62	0.80	0.87	0.96	0.99	1.00	1.00
SM-LG-SM	K1	0.59	0.78	0.85	0.95	0.99	1.00	1.00
NUMBER17	K1	0.41	0.70	0.82	0.95	1.00	1.00	1.00
X000	K1	0.52	0.69	0.78	0.90	0.97	0.99	0.99
NUMBER23	K1	0.33	0.62	0.76	0.93	0.99	1.00	1.00
3RD LINE	K1	0.31	09.0	0.74	0.92	0.99	1.00	1.00
3+2 CARS	K1	0.34	0.58	0.70	0.87	0.97	0.99	1.00
HALFOVAL	K1	0.51	0.65	0.73	0.85	0.95	0.97	0.98
$_{-}78910$	K1	0.30	0.59	0.73	0.91	0.99	1.00	1.00
2+3STICK	K1	0.30	0.55	0.68	0.87	0.98	0.99	1.00
#BUGS	K1	0.44	0.63	0.74	0.89	0.98	0.99	1.00
2 + 2	K1	0.23	0.53	0.70	0.92	1.00	1.00	1.00
3 + 3	K1	0.16	0.47	0.66	0.91	1.00	1.00	1.00
1 + 7	K1	0.24	0.47	0.60	0.82	0.96	0.98	0.99
TEAMS_R	3	0.38	0.55	0.64	0.81	0.94	0.97	0.98
VICKS R	3	0.15	0.40	0.57	0.85	0.98	1.00	1.00
See notes at end of table.								

Table C2. Mathematics assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, 2003–04, and 2006–07

999–2000, 2001–02, 2003–04, and	
Mathematics assessment estimated proportion correct: School years 1998–99, 1	2006–07—Continued
Table C2.	

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
8-6CRAYN	Kl	0.21	0.41	0.54	0.77	0.93	0.97	0.99
3 + 4	K1	0.12	0.34	0.51	0.80	0.97	0.99	1.00
5-10RANG	K1	0.23	0.41	0.55	0.80	0.97	0.99	1.00
2+5MARBL	K1,3	0.17	0.35	0.48	0.73	0.92	0.97	0.98
SHAPES	K1	0.41	0.51	0.58	0.70	0.83	0.89	0.92
PATTERN	K1	0.33	0.47	0.57	0.76	0.93	0.97	0.99
2+5CIRCL	K1	0.12	0.30	0.44	0.71	0.93	0.97	0.99
12 BY 2S	K1,3	0.09	0.27	0.42	0.73	0.95	0.98	0.99
3+7PENNY	K1,3	0.08	0.24	0.39	0.70	0.94	0.98	0.99
51015_25	K1,3	0.05	0.19	0.33	0.66	0.93	0.98	0.99
ORANGE_R	33	0.21	0.33	0.44	0.68	0.91	0.97	0.99
11 + 3	K1	0.05	0.18	0.31	0.64	0.93	0.98	0.99
7 - 3	K1	0.04	0.15	0.28	0.63	0.94	0.98	1.00
9 - 2	K1	0.03	0.14	0.27	0.62	0.94	0.98	1.00
PATHS_R	33	0.12	0.25	0.34	0.56	0.81	0.90	0.94
6+7	K1	0.03	0.13	0.24	0.56	0.89	0.96	0.99
12 + 6	K1	0.04	0.12	0.22	0.49	0.84	0.93	0.97
# MORE	K1	0.03	0.11	0.20	0.47	0.84	0.93	0.97
MOST_Y	33	0.02	0.08	0.17	0.47	0.87	0.96	0.99
2-1+2	K1	0.04	0.12	0.20	0.46	0.81	0.92	0.97
RULER_R	3	0.06	0.15	0.23	0.44	0.74	0.86	0.92
$A13_79$	K1,3,5	0.03	0.10	0.18	0.42	0.79	0.91	0.96
SIDES_R	ŝ	0.14	0.20	0.27	0.48	0.79	0.90	0.95
4+4-2	K1,3	0.02	0.07	0.14	0.40	0.81	0.92	0.97
PAGES_R	3	0.21	0.25	0.31	0.51	0.84	0.94	0.98
17 - 4	K1	0.01	0.06	0.12	0.38	0.81	0.93	0.98
$COST_{10}$	K1,3,5	0.02	0.07	0.14	0.38	0.80	0.92	0.97
12-9	K1	0.01	0.05	0.12	0.37	0.80	0.93	0.98
26 + 20	K1	0.01	0.04	0.10	0.33	0.78	0.92	0.97
CARS15_5	K1,3,5	0.01	0.05	0.11	0.33	0.76	0.91	0.97
FEWESTY	3	0.01	0.04	0.09	0.30	0.75	0.90	0.96
See notes at end of table.								

2. Mathematics assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, 2003–04, and	2006–07—Continued
Table C	

	IIsed in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
SOUARE R	ŝ	0.08	0.16	0.23	0.38	0.64	0.76	0.84
CUBES10	3,5	0.08	0.16	0.22	0.38	0.63	0.75	0.84
HOWMANY\$	K1,3	0.02	0.07	0.12	0.31	0.67	0.83	0.91
CANDY8_2	K1,3,5	0.01	0.03	0.08	0.26	0.70	0.87	0.95
BEADS_R	ŝ	0.00	0.01	0.04	0.19	0.71	0.90	0.97
12-? PEN	K1,3	0.01	0.03	0.06	0.23	0.68	0.86	0.95
NEXT78	3,5	0.01	0.03	0.06	0.23	0.68	0.86	0.95
HEADSUP	K1,3	0.04	0.10	0.15	0.31	0.61	0.76	0.86
24-14BKS	K1	0.00	0.02	0.05	0.21	0.67	0.86	0.95
MEANS_R	ω	0.00	0.02	0.04	0.18	0.64	0.85	0.94
EQUAL_R	ω	0.17	0.18	0.20	0.31	0.70	0.88	0.95
D0_ADD4	3,5	0.01	0.03	0.06	0.21	0.63	0.82	0.92
MONEY_R	ω	0.00	0.01	0.03	0.16	0.65	0.86	0.95
TIME1030	3,5	0.01	0.03	0.07	0.21	0.61	0.81	0.91
POINTS_R	ω	0.30	0.32	0.34	0.44	0.71	0.85	0.93
SCORE_Y	ξ	0.00	0.01	0.03	0.15	0.61	0.83	0.94
GOALS	K1,3	0.00	0.02	0.05	0.18	0.59	0.80	0.91
PAPER	ς	0.00	0.01	0.03	0.15	0.59	0.82	0.93
NICKELS	ŝ	0.00	0.02	0.04	0.16	0.57	0.79	0.91
17CENTS	K1	0.00	0.01	0.03	0.13	0.56	0.79	0.91
MORE1_Y	ω	0.00	0.01	0.02	0.10	0.55	0.80	0.92
NUMBER60	3,5	0.00	0.01	0.02	0.10	0.54	0.80	0.92
BDCAKE	K1	0.00	0.02	0.04	0.14	0.53	0.76	0.88
CUBESIDE	3,5	0.02	0.05	0.09	0.21	0.51	0.69	0.81
NEXT120	3,5	0.00	0.01	0.02	0.10	0.50	0.76	0.89
FEWER_Y	3	0.00	0.00	0.01	0.08	0.50	0.77	0.91
TREES -MC	8	0.10	0.11	0.12	0.19	0.55	0.77	06.0
CHART_64	3,5	0.00	0.02	0.04	0.14	0.49	0.72	0.85
AGEGRAPH	5	0.01	0.03	0.05	0.16	0.48	0.69	0.83
BOX_700	3,5	0.00	0.00	0.01	0.07	0.48	0.75	0.89
NUMBER	3	0.00	0.01	0.02	0.10	0.47	0.72	0.87
See notes at end of table.								

ars 1998–99, 1999–2000, 2001–02, 2003–04, and	
22. Mathematics assessment estimated proportion correct: School ye	2006–07—Continued
Table C	

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
BSP70Z	3,5,8	0.00	0.01	0.02	0.10	0.46	0.72	0.86
ANDY27	5	0.00	0.01	0.01	0.08	0.46	0.73	0.87
REES100	5	0.00	0.01	0.02	0.08	0.45	0.71	0.87
OLORSYM	3,5	0.00	0.01	0.03	0.12	0.45	0.69	0.83
RIES	3	0.00	0.01	0.02	0.08	0.45	0.71	0.86
CHILDR_Y	3	0.00	0.01	0.02	0.08	0.44	0.70	0.86
TAR-Y	3	0.02	0.04	0.07	0.17	0.45	0.64	0.77
GSLEFT	3,5,8	0.08	0.08	0.09	0.15	0.47	0.71	0.86
30XSHELF	5	0.00	0.02	0.04	0.12	0.43	0.66	0.81
SECOND_Y	3	0.00	0.01	0.02	0.08	0.42	0.68	0.84
SAMJUAN	8	0.11	0.12	0.12	0.15	0.47	0.73	0.88
A568214K	3,5	0.00	0.01	0.01	0.07	0.41	0.68	0.84
IST#X5	5	0.22	0.23	0.24	0.30	0.55	0.73	0.86
BIKETIME	5	0.00	0.01	0.02	0.08	0.40	0.65	0.81
MINSSIM	5,8	0.01	0.03	0.05	0.13	0.41	0.61	0.76
FRUIT	3	0.12	0.13	0.14	0.20	0.46	0.67	0.81
24/4 TAB	K1	0.01	0.02	0.04	0.11	0.39	0.61	0.76
SCALE=	5	0.00	0.01	0.02	0.08	0.36	0.60	0.77
CARWASH	3,5,8	0.00	0.01	0.02	0.07	0.34	0.59	0.77
SHANBIKE	8	0.10	0.11	0.12	0.17	0.41	0.62	0.77
FLIPTRI	8	0.14	0.16	0.18	0.25	0.46	0.61	0.74
SCALEMAR	8	0.12	0.13	0.14	0.20	0.42	0.61	0.76
CARDS579	3,5	0.00	0.00	0.01	0.05	0.31	0.57	0.75
MARIA	3,5,8	0.00	0.01	0.01	0.06	0.32	0.56	0.75
24LEMONS	5	0.00	0.00	0.01	0.04	0.29	0.55	0.74
<i>TILESCOV</i>	3	0.01	0.02	0.03	0.09	0.33	0.53	0.70
PROBNOT	8	0.16	0.16	0.17	0.21	0.41	0.61	0.76
100MORE	3,5,8	0.10	0.10	0.10	0.13	0.34	0.57	0.75
AREA_B	3	0.00	0.01	0.02	0.06	0.29	0.51	0.69
CARGER_B	3	0.00	0.01	0.01	0.05	0.26	0.49	0.68
LOWDIVE	8	0.14	0.14	0.15	0.18	0.36	0.56	0.72
See notes at end of table.								

999–2000, 2001–02, 2003–04, and	
Mathematics assessment estimated proportion correct: School years 1998–99, 1	2006–07—Continued
Table C2.	

	Used in Grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
FOURLINE	8	0.19	0.20	0.21	0.25	0.42	0.57	0.71
TITO	3,5,8	0.00	0.01	0.01	0.04	0.24	0.47	0.66
PENCIL_Y	3	0.06	0.07	0.09	0.15	0.34	0.50	0.64
LOUISA13	3,5	0.00	0.00	0.00	0.02	0.20	0.46	0.68
KYLEBOB	5,8	0.11	0.12	0.12	0.15	0.33	0.52	0.69
OPOS-MC	8	0.28	0.28	0.29	0.30	0.43	0.61	0.76
EQUAL_B	3	0.10	0.10	0.10	0.13	0.30	0.50	0.68
AGE1/4	5	0.24	0.24	0.24	0.24	0.37	0.57	0.75
1/4STU44	5	0.00	0.00	0.00	0.01	0.15	0.43	0.67
LONGSTEP	5	0.00	0.00	0.01	0.03	0.19	0.40	0.60
7XPLUS4	8	0.19	0.20	0.20	0.22	0.35	0.52	0.68
NUMBE2_B	3	0.00	0.00	0.00	0.02	0.17	0.39	09.0
BEADSWHT	5	0.00	0.00	0.01	0.02	0.16	0.37	0.58
НОЦЦУ	3,5,8	0.00	0.00	0.01	0.03	0.17	0.37	0.57
$TALL75_{-}$	3,5	0.10	0.10	0.10	0.11	0.22	0.42	0.63
X8ANDYX	8	0.25	0.25	0.25	0.27	0.38	0.52	0.67
YDOLLARS	8	0.09	0.09	0.09	0.12	0.25	0.42	0.59
CHANGE	K1	0.00	0.00	0.01	0.03	0.17	0.36	0.56
MARBLES	3,5	0.00	0.00	0.00	0.01	0.12	0.33	0.56
BANKER	3,5	0.00	0.00	0.01	0.03	0.16	0.35	0.55
SIDDORNG	8	0.22	0.22	0.22	0.22	0.31	0.46	0.64
MYSTER_B	3	0.00	0.00	0.00	0.01	0.11	0.30	0.53
NUMEDGES	3,5,8	0.00	0.01	0.02	0.05	0.19	0.35	0.51
OJ 300Z	5	0.00	0.00	0.00	0.01	0.10	0.29	0.52
FRAME3FT	5	0.00	0.00	0.00	0.01	0.10	0.29	0.51
SAND40MC	5,8	0.00	0.00	0.00	0.00	0.06	0.23	0.47
SAMEFRAC	3,5	0.00	0.00	0.00	0.01	0.09	0.25	0.45
MRKRULER	3,5,8	0.00	0.00	0.00	0.01	0.10	0.25	0.45
NUMSENT5	8	0.13	0.13	0.13	0.13	0.19	0.32	0.51
SMITHFAM	5,8	0.21	0.21	0.21	0.21	0.25	0.36	0.54
SHADED.2	5	0.11	0.11	0.11	0.11	0.16	0.29	0.48
See notes at end of table.								

and	
2003–04,	
-02,	
2001-	
00, 2	
9-20	
, 199	
8–99	
199	
years	
loor	
t: Scl	
orrec	
on co	
porti	
d pro	
mate	
t estin	
ment	led
ssess	ntinu
ics a	ç
emat	-0-1
Math	2006
5.	
ble C	
Та	

Round 7	0.49	0.46	0.41	0.42	0.38	0.53	0.45	0.33	0.29	0.37	0.30	0.38	0.38	0.25	0.38	0.18	0.22	0.11	0.22	me are corted in
Round 6	0.31	0.27	0.23	0.23	0.17	0.41	0.28	0.15	0.10	0.20	0.12	0.23	0.25	0.10	0.27	0.05	0.12	0.04	0.16	t comparisons Ite
Round 5	0.20	0.15	0.10	0.09	0.04	0.31	0.18	0.04	0.01	0.12	0.03	0.14	0.18	0.02	0.20	0.01	0.07	0.01	0.13	non scale to suppor
Round 4	0.15	0.11	0.01	0.04	0.00	0.24	0.15	0.00	0.00	0.09	0.00	0.12	0.16	0.00	0.15	0.00	0.05	0.00	0.12	heen nut on a comr
Round 3	0.15	0.11	0.00	0.04	0.00	0.22	0.15	0.00	0.00	0.09	0.00	0.12	0.16	0.00	0.15	0.00	0.05	0.00	0.12	eved about and have
Round 2	0.15	0.11	0.00	0.04	0.00	0.21	0.15	0.00	0.00	0.09	0.00	0.12	0.16	0.00	0.14	0.00	0.05	0.00	0.12	r bindercerten through
Round 1	0.15	0.11	0.00	0.04	0.00	0.21	0.15	0.00	0.00	0.09	0.00	0.12	0.16	0.00	0.14	0.00	0.05	0.00	0.12	round Estimates fo
Used in Grades	8	5,8	8	5,8	5	5	5,8	3	5	5	5,8	8	8	5	8	5,8	8	3,5,8	8	in correct for each item in each
	X5PLUS3X	3FRAC4	LIFEEXP	RECPLAY	OPOSITIV	SALESTAX	PIZZA	FENCEB	SHADED.3	DIFF=88	VOLPRISM	500BATT	CROSSLIN	MEASDIAM	NPLUS5	CARPSFOE	CIRCCIRC	12INCH	FOLDCUBE	NOTE IBT estimated monstrip

For the inter-exerting order of overall difficulty (IRT "b" parameter). The grades in which items appeared on assessment forms are over put on a common scate to support comparisons. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). The grades in which items appeared on assessment forms are noted. Not all items appeared in test forms for all rounds. Items whose format changed from multiple choice in grade 5 to open ended in grade 8 were not treated as common items, and are listed separately. Table estimates are based on cross sectional-weights within each round (CICW0, C3CW0, C4CW0, C5CW0, C6CW0, C7CW0). SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2004, and spring 2007.

Table C3. Science	assessment estimated pro	portion correct: Schoo	l years 2001–02, 2003-	-04, and 2006–07
	Used in Grades	Round 5	Round 6	Round 7
RBULB	3	06.0	0.95	96.0
RENRGY	3	0.94	0.98	0.99
RPLANT	3	0.94	0.98	1.00
RORGAN	3	0.79	0.86	0.93
RTOOL	3	0.84	0.92	0.97
ROUIMM	3,5	0.84	0.92	0.98
RDSAST	3	0.88	0.96	0.99
RFGRPS	3	0.78	0.86	0.93
RFORMS	3	0.82	0.91	0.97
YPLAIN	3	0.73	0.84	0.93
RWINGS	3,5	0.82	0.93	0.98
RANIML	3	0.75	0.87	0.95
ROUFRZ	3,5	0.77	0.90	0.97
ROCCUR	3	0.79	0.91	0.97
WHCHPREY	5	0.69	0.85	0.95
RSEEDS	3	0.67	0.81	0.92
ROUTAP	3,5	0.60	0.72	0.85
ROUJUN	3,5	0.67	0.84	0.95
RTHING	3	0.72	0.85	0.95
RWATER	3	0.67	0.82	0.93
YDSAST	3	0.66	0.79	0.90
RSUNIS	3	0.73	0.86	0.95
SOLAR	8	0.65	0.79	0.91
ROUBRN	3,5	0.62	0.81	0.94
ROUERT	3,5,8	0.63	0.77	06.0
RFISHB	3	0.63	0.79	0.92
RSHAPE	3	0.67	0.81	0.93
RHEART	3,5	0.59	0.78	0.93
RPWDER	3	0.63	0.78	0.91
ROUJAR	3,5	0.55	0.68	0.83
CUTSCAB	5,8	0.60	0.76	0.90
See notes at end of table.				

, and 2006–(
12, 2003-04
rears 2001–(
ct: School y
ortion corre
imated prop
sessment est
Science as
Table C3.

ROUSRF 3,5 0.72 0.84 0.90 PYRAMID 5,8 0.53 0.67 0.82 PYRAMID 5,8 0.53 0.67 0.83 YTHENT 3,5,8 0.51 0.76 0.83 YTHENT 3,5,8 0.54 0.66 0.83 YTHENT 3,5,8 0.54 0.66 0.84 BEARTH 3,5,8 0.54 0.66 0.84 YSOUND 3 0.49 0.65 0.83 YSOUND 3,5,8 0.41 0.66 0.86 YNOON 3,5,8 0.41 0.61 0.86 YNOON 3,5,8 0.41 0.61 0.86 YNOON 3,5,8 0.41 0.61 0.86 YNOON 3,5,8 0.43 0.66 0.86 YNOON 3,5 0.49 0.66 0.86 YNOON 3,5 0.44 0.61 0.86 ROUSOL 3,5 0.48		Used in Grades	Round 5	Round 6	Round 7
RDESRT 3,5 0.60 0.76 0.93 PYRAMID 5,8 0.53 0.67 0.82 WTNSNOW 5,8 0.51 0.70 0.83 YTHENT 3,5,8 0.54 0.67 0.84 YTHENT 3,5,8 0.54 0.66 0.84 WINSNOW 5,8 0.54 0.66 0.84 WINSCT 3,5,8 0.54 0.66 0.86 YNOND 3,5,8 0.56 0.73 0.86 YNNSCT 3,5,8 0.56 0.73 0.86 YNNSCT 3,5,8 0.56 0.77 0.86 YNNSCT 3,5,8 0.44 0.66 0.86 YNNOON 5,8 0.47 0.66 0.86 YNNOON 5,8 0.47 0.66 0.86 ROUND 3,5 0.44 0.66 0.86 ROUSOL 3,5 0.44 0.66 0.86 ROUSOL 3,5 0.48	ROUSRF	3,5	0.72	0.84	0.94
PYRAMID 5,8 0.53 0.67 0.83 WTHEMIT 3,5,8 0,51 0,70 0,88 YTHEMIT 3,5,8 0,51 0,65 0,88 YTHEMIT 3,5,8 0,51 0,65 0,88 YSOUND 3 0,49 0,65 0,88 YSOUND 3 0,56 0,73 0,90 YNNSCT 3,5,8 0,59 0,77 0,86 YNNSCT 3,5,8 0,59 0,77 0,80 YNNSCT 3,5,8 0,41 0,61 0,83 YSENSE 3 0,41 0,61 0,83 RAVMON 5,8 0,47 0,66 0,86 ROUNSOL 3,5 0,47 0,66 0,86 RAVMON 5,8 0,47 0,66 0,86 ROUSOL 3,5 0,48 0,67 0,86 NATRGRPH 5,8 0,46 0,67 0,86 NARPOLL 5,8 0,46	RDESRT	3,5	0.60	0.76	0.90
MTNSNOW 5 0.51 0.70 0.88 YTHEMT 3,5,8 0,54 0,68 0,84 YERNETH 3 0,49 0,67 0,84 SUEARTH 3 0,51 0,67 0,84 SUEARTH 3 0,51 0,67 0,86 YSOUND 3,5 0,51 0,67 0,86 YNNSCT 3 0,51 0,67 0,86 YNOON 3,5,8 0,59 0,75 0,90 YNOON 3,5,8 0,59 0,75 0,86 YSENSE 3 0,41 0,66 0,86 PROTECT 5 0,47 0,66 0,86 RAVMON 5,8 0,47 0,66 0,86 ROUSOL 3,5 0,48 0,67 0,86 ROUSOL 3,5 0,34 0,59 0,86 ROUSOL 3,5 0,34 0,56 0,86 ROUSOL 3,5 0,34 0,56	PYRAMID	5,8	0.53	0.67	0.82
YTHEMT 3,5,8 0.54 0.68 0.81 BEARTH 3 0.49 0.65 0.83 SUGARDIS 5 0.49 0.65 0.83 SUGARDIS 3 0.51 0.66 0.83 YNSCT 3 0.51 0.67 0.86 0.80 YNSCT 3,5,8 0.59 0.73 0.90 0.86 0.80 YNSCT 3,5,8 0.50 0.73 0.90 0.83 0.90 0.86 0.80 YNDON 3,5,8 0.50 0.71 0.61 0.83 0.91 YSENSE 3 0.41 0.61 0.66 0.88 0.86 YNDON 5,8 0.44 0.66 0.86 0.86 ROUSOL 3,5 0.34 0.65 0.86 0.86 ROUSOL 3,5 0.34 0.65 0.86 0.86 ROUSOL 3,5 0.34 0.65 0.86 0.86 ROUNDL	MONSNTM	5	0.51	0.70	0.88
BEARTH 3 0.49 0.65 0.83 YSOUND 3 0.51 0.67 0.86 YSOUND 3 0.51 0.67 0.86 YNON 3,5,8 0.59 0.73 0.90 YNON 3,5,8 0.59 0.75 0.90 YNON 3,5,8 0.59 0.75 0.90 YSENSE 3 0.41 0.61 0.83 PROTECT 5 0.47 0.66 0.86 PROTECT 5 0.47 0.66 0.86 PROTECT 3,5 0.48 0.66 0.86 PROTECT 3,5 0.48 0.66 0.86 RAVMOON 5,8 0.48 0.66 0.86 ROUSOL 3,5 0.48 0.66 0.86 ROUGRT 5,8 0.46 0.66 0.86 ROUGRT 3,5 0.34 0.55 0.80 ROUGRT 3,5 0.34 0.55	YTHEMT	3,5,8	0.54	0.68	0.84
SUGARDIS 5 0.48 0.67 0.86 YSOUND 3 0.51 0.68 0.86 YNOON 3,5,8 0.51 0.68 0.86 YNOON 3,5,8 0.55 0.73 0.90 YNOON 3,5,8 0.59 0.75 0.91 YENSE 3 0.41 0.61 0.83 FARTHOK 5,8 0.43 0.61 0.83 FANDON 3,5,8 0.47 0.66 0.86 PROTECT 5 0.47 0.66 0.86 PROTECT 3,5 0.48 0.66 0.86 ROUSOL 3,5 0.48 0.66 0.86 ROUSOL 3,5 0.34 0.55 0.80 ROUSOL 3,5 0.34 0.55 0.80 ROUGRT 3,5 0.34 0.55 0.80 ROUGRT 3,5 0.34 0.55 0.80 ROUMC 3,5 0.34 0.55 </td <td>BEARTH</td> <td>ŝ</td> <td>0.49</td> <td>0.65</td> <td>0.83</td>	BEARTH	ŝ	0.49	0.65	0.83
YSOUND 3 0.51 0.68 0.80 YINSCT 3 0.56 0.73 0.90 YNSCT 3 0.59 0.75 0.91 YNSCR 3 0.59 0.75 0.91 YSENSE 3 0.41 0.61 0.83 YSENSE 0.53 0.47 0.66 0.83 FAVMOON 5,8 0.47 0.66 0.86 PROTECT 3,5 0.47 0.66 0.86 PROTECT 3,5 0.47 0.66 0.86 BSHADW 3,5 0.47 0.66 0.86 ROUSOL 3,5 0.34 0.67 0.81 WATRGRPH 5,8 0.46 0.61 0.86 WATRGRPH 5,8 0.46 0.66 0.81 WATRGRPH 5,8 0.46 0.66 0.86 ROUGRT 3,5 0.34 0.50 0.81 ROUGRT 3,5 0.34 0.56 <td>SUGARDIS</td> <td>5</td> <td>0.48</td> <td>0.67</td> <td>0.86</td>	SUGARDIS	5	0.48	0.67	0.86
YINSCT 3 0.56 0.73 0.90 YMOON 3,5,8 0.59 0.75 0.91 YSENSE 3 0.41 0.61 0.83 YSENSE 3 0.41 0.61 0.83 YSENSE 3,5,8 0.53 0.72 0.89 FARTHQK 5,8 0.43 0.66 0.85 GRAVMOON 5,8 0.47 0.66 0.86 BSHADW 3,5 0.47 0.66 0.86 BSHADW 3,5 0.47 0.66 0.86 BSHADW 3,5 0.48 0.66 0.86 ROUSCL 3,5 0.44 0.66 0.86 MATRGRPH 5,8 0.46 0.61 0.81 WATRGRPH 5,8 0.46 0.66 0.86 WATRGRPH 3,5 0.34 0.50 0.81 ROUBLB 3,5 0.46 0.61 0.81 ROUMCH 3,5 0.34	V SOUND	3	0.51	0.68	0.86
YMOON3,5,80.590.750.91YSENSE30.410.610.83YSENSE5,80.530.720.89GRAVMOON5,80.470.660.86PROTECT50.470.660.86PROTECT3,50.470.660.86PROTECT3,50.470.660.86PROTECT3,50.470.660.86PROTECT3,50.480.660.86PROTECT3,50.480.670.86PROTECT3,50.340.670.84AIRPOLL5,80.340.670.81AIRPOLL5,80.340.670.81MAPOLL5,80.340.670.81NUDER3,50.340.670.81ROUGRT3,50.340.670.81ROUGRT3,50.340.620.81ROUGRT3,50.340.650.81ROUGRT3,50.340.650.86ROUMC3,50.340.650.86ROUMC3,50.340.650.86ROUMC3,50.340.650.81YDSOLV30.310.560.75YDSOLV3,50.300.410.50YDSOLV3,50.300.470.77YDSOLV3,50.300.470.77	YINSCT	3	0.56	0.73	0.90
YSENSE 3 0.41 0.61 0.83 EARTHQK 5,8 0.53 0.72 0.89 EARTHQK 5,8 0.47 0.66 0.86 PROTECT 5 0.47 0.66 0.86 PROTECT 5 0.47 0.66 0.86 PROTECT 3,5 0.48 0.66 0.86 BSHADW 3,5 0.48 0.66 0.86 BSHADW 3,5 0.48 0.67 0.80 ROUSOL 3,5 0.34 0.59 0.84 AIRPOLL 5,8 0.46 0.61 0.80 AIRPOLL 5,8 0.46 0.67 0.80 WATRGRPH 5,8 0.46 0.66 0.80 NUBELS 3,5 0.34 0.55 0.80 ROUBLB 3,5 0.34 0.56 0.80 ROUMTN 3,5 0.34 0.56 0.80 ROUMCE 3,5 0.38 0.5	YMOON	3,5,8	0.59	0.75	0.91
EARTHQK5,80.530.720.89RAVMOON5,80.470.660.86PROTECT50.470.660.86BSHADW30.480.660.86BSHADW3,50.470.660.86BSHADW3,50.500.660.86BSHADW3,50.500.660.86BSHADW3,50.500.670.86ROUSOL3,50.340.670.84NDDLL5,80.340.670.81WATRGRPH5,80.460.610.81VBEES3,50.340.590.81NOUBLB3,50.340.610.81ROUGRT3,50.340.550.80ROUGRT3,50.340.560.81ROUGRT3,50.340.560.81ROUMTN3,5,80.340.550.80ROUNCE3,50.340.560.86ROUNCE3,50.340.560.86ROUNCE3,50.340.560.86ROUNCE3,50.340.560.75BSOUND3,5,80.340.560.75YDSOLV30.340.560.75YDSOLV3,50.310.560.75YDSOLV3,50.310.500.77YDSOLV3,50.300.470.70YDSOLV3,50.300.770.77YDSO	YSENSE	3	0.41	0.61	0.83
GRAVMOON5,80.480.660.86PROTECT50.470.660.86BSHADW30.480.660.85BSHADW3,50.480.660.85BSHADW3,50.500.660.86ROUSOL3,50.500.660.86ROUSOL3,50.2480.670.86ROUSOL5,80.480.670.86ROUSOL5,80.340.670.86AIRPOLL50.340.590.81WATRGRPH5,80.460.610.81VBEES3,50.340.550.80ROUBLB3,50.480.610.81ROUGRT3,50.480.620.81ROUGRT3,50.480.640.86ROUMTN3,50.340.560.86ROUMTN3,50.460.640.85ROUNCE3,50.380.560.75BSOUND3,50.380.580.75VDSOLV3,50.310.560.75VDSOLV3,50.310.500.76ROUSHD3,50.310.500.77ROUSHD3,50.310.500.77ROUSHD3,50.310.500.77ROUSHD3,50.310.770.77ROUSHD3,50.310.770.77ROUSHD3,50.310.770.77ROUSH	EARTHQK	5,8	0.53	0.72	0.89
PROTECT 5 0.47 0.66 0.86 BSHADW 3,5 0.48 0.66 0.85 BSHADW 3,5 0.48 0.66 0.85 ROUSOL 3,5 0.48 0.66 0.85 THUNDER 5,8 0.48 0.66 0.82 AIRPOLL 5 0.34 0.59 0.80 MRPOLL 5,8 0.46 0.61 0.81 WATKGRPH 5,8 0.34 0.59 0.80 WATRGRPH 5,8 0.46 0.61 0.81 WATRGRPH 5,8 0.46 0.61 0.81 ROUBLB 3,5 0.34 0.55 0.80 ROUBRD 3,5 0.34 0.56 0.81 ROUMTN 3,5 0.46 0.64 0.86 ROUMTN 3,5 0.38 0.56 0.80 ROUMTN 3,5 0.38 0.56 0.75 ROUNCE 3,5 0.38 0	GRAVMOON	5,8	0.48	0.66	0.86
BSHADW30.480.660.85ROUSOL3,50.500.660.82THUNDER5,80.480.670.86AIRPOLL50.340.670.81WATRGRPH5,80.460.610.81WATRGRPH5,80.460.610.81WATRGRPH5,80.460.610.81WATRGRPH5,80.460.610.81WATRGRPH3,50.340.550.80ROUBLB3,50.480.640.81ROUGRT3,50.480.640.80ROUMTN3,5,80.460.640.86ROUMTN3,5,80.460.640.82ROUMTN3,5,80.460.640.87ROUMTN3,5,80.340.560.75ROUMTN3,5,80.340.560.75ROUMTN3,5,80.340.560.75ROUND3,5,80.340.640.87ROUND3,5,80.340.560.75POUND3,5,80.340.560.75YDSOLV3,5,80.340.560.75YDSOLV3,50.310.500.75ROUSHD3,50.300.470.70ROUSHD3,50.300.470.77	PROTECT	5	0.47	0.66	0.86
ROUSOL3,50.500.650.82THUNDER5,80.480.670.84AIRPOLL5,80.480.670.84MATROLL5,80.340.590.84WATRGRPH5,80.460.610.81WATRGRPH3,50.340.550.81VBEES3,50.340.550.81VBEES3,50.340.650.81NDELB3,50.340.550.81ROUBLB3,50.370.620.81ROUGRT3,50.370.640.80ROUMTN3,50.460.640.85ROUMTN3,50.380.460.64ROUMTN3,50.380.580.75ROUND3,50.380.560.75NDSOLV3,50.310.560.75VDSOLV3,50.310.500.75ROUSHD3,50.300.470.70	BSHADW	ŝ	0.48	0.66	0.85
THUNDER5,80.480.670.86AIRPOLL50.340.590.84WATRGRPH5,80.340.590.81WATRGRPH5,80.340.610.81WATRGRPH3,50.340.550.80NBEES3,50.340.550.80ROUBLB3,50.480.620.81ROUGRT3,5,80.460.640.86ROUMTN3,5,80.460.640.85ROUMCE3,50.380.560.82ROUMCE3,50.380.580.75BOUND3,5,80.380.580.75BSOUND3,5,80.380.540.75VDSOLV3,50.310.560.75ROUSHD3,50.310.500.75ROUSHD3,50.300.470.70	ROUSOL	3,5	0.50	0.65	0.82
AIRPOLL50.340.590.84WATRGRPH5,80.460.610.81WATRGRPH5,80.460.610.81YBEES3,50.340.550.80ROUBLB3,50.370.560.80ROUGRT3,50.370.620.80ROUMTN3,5,80.460.640.80ROUMTN3,5,80.460.640.85ROUMCE3,50.380.560.75BOUND3,5,80.380.580.75BSOUND3,5,80.380.560.75VDSOLV30.310.560.75ROUSHD3,50.300.410.560.75	THUNDER	5,8	0.48	0.67	0.86
WATRGRPH5,80.460.610.81YBEES3,50.340.550.80YBEES3,50.340.550.80ROUBLB3,50.370.560.80ROUGRT3,5,80.450.640.80ROUMTN3,5,80.460.640.86ROUMTN3,5,80.460.640.86ROUMTN3,5,80.460.640.85ROUMCE3,50.380.580.82ROUFLY3,5,80.380.580.75BOUND3,5,80.380.560.75VDSOLV3,5,80.310.560.76VDSOLV3,50.310.560.76ROUSHD3,50.300.470.75	AIRPOLL	5	0.34	0.59	0.84
YBEES3,50.340.550.80ROUBLB3,50.480.620.81ROUGRT3,50.370.620.81ROUGRT3,5,80.370.560.80ROUMTN3,5,80.450.640.86ROUMTN3,5,80.460.640.85ROUMTN3,50.380.580.82ROUMCE3,50.380.580.75BOUND3,5,80.380.580.75BSOUND3,5,80.310.560.75VDSOLV3,50.310.500.75ROUSHD3,50.300.470.70	WATRGRPH	5,8	0.46	0.61	0.81
ROUBLB3,50.480.620.81ROUGRT3,50.370.560.80ROUMTN3,5,80.370.560.80ROUMTN3,5,80.450.640.85ROUMTV3,50.460.640.85ROUMCE3,50.380.580.75BOUNTLY3,5,80.380.580.75BSOUND3,5,80.380.580.75VDSOLV3,5,80.310.560.76VDSOLV3,50.310.500.75ROUSHD3,50.300.470.70	YBEES	3,5	0.34	0.55	0.80
ROUGRT3,50.370.560.80ROUMTN3,5,80.450.640.86ROUMTN3,5,80.460.640.85ROUMCE3,50.380.580.82ROUFLY3,50.380.580.75BOUND3,5,80.380.580.75VDSOLV3,5,80.380.560.75VDSOLV3,5,80.310.560.76VDSOLV3,50.310.500.76ROUSHD3,50.300.470.70	ROUBLB	3,5	0.48	0.62	0.81
ROUMTN3,5,80.450.640.86ROUMCE3,50.460.640.85ROUMCE3,50.350.580.82ROUFLY3,50.380.580.82BOUND3,5,80.460.580.75BSOUND3,5,80.380.540.77YDSOLV30.410.560.76LAMPWIRE50.310.500.75ROUSHD3,50.300.470.75	ROUGRT	3,5	0.37	0.56	0.80
ROUMCE3,50.460.640.85ROUFLY3,50.380.580.82ROUFLY3,50.380.580.75ECLIPSE880.460.580.75BSOUND3,5,80.380.540.77YDSOLV30.410.560.76LAMPWIRE50.310.500.75ROUSHD3,50.300.470.70	ROUMTN	3,5,8	0.45	0.64	0.86
ROUFLY3,50.380.580.82ECLIPSE880.460.580.75BSOUND3,5,80.380.540.77YDSOLV30.410.560.76LAMPWIRE50.310.500.75ROUSHD3,50.300.470.70	ROUMCE	3,5	0.46	0.64	0.85
ECLIPSE880.460.580.75BSOUND3,5,80.380.540.77YDSOLV30.410.560.76VDSOLV50.310.500.75LAMPWIRE50.300.470.72	ROUFLY	3,5	0.38	0.58	0.82
BSOUND 3,5,8 0.38 0.54 0.77 YDSOLV 3 0.41 0.56 0.76 YDSOLV 3 0.41 0.56 0.76 LAMPWIRE 5 0.31 0.50 0.75 ROUSHD 3,5 0.30 0.47 0.72	ECLIPSE8	8	0.46	0.58	0.75
YDSOLV 3 0.41 0.56 0.76 LAMPWIRE 5 0.31 0.50 0.75 ROUSHD 3,5 0.30 0.47 0.72	BSOUND	3,5,8	0.38	0.54	0.77
LAMPWIRE 5 0.31 0.50 0.75 ROUSHD 3,5 0.30 0.47 0.72	YDSOLV	3	0.41	0.56	0.76
ROUSHD 3,5 0.30 0.47 0.72	LAMPWIRE	5	0.31	0.50	0.75
	ROUSHD	3,5	0.30	0.47	0.72

Science assessment estimated proportion correct: School years 2001–02, 2003–04, and 2006–07—Continued Table C3.

	Used in Grades	Round 5	Round 6	Round 7
MIXTURE	5	0.32	0.51	0.79
ECLIPSE5	5	0.26	0.46	0.74
YFWATE	3	0.42	0.59	0.82
BPLNT2	3,5	0.34	0.51	0.75
YLIVE	3	0.29	0.48	0.77
CUPTEMP	5	0.22	0.41	0.71
BPLANT	3,5	0.32	0.48	0.75
YBLANC	3	0.23	0.41	0.70
BHIBER	3	0.42	0.54	0.73
YFARMG	3	0.32	0.43	0.61
EARTHCTR	8	0.42	0.54	0.74
BURIED	5,8	0.36	0.48	0.68
LIMESTON	8	0.39	0.52	0.74
BSLIDE	3,5,8	0.26	0.39	0.63
SEEDGROW	5,8	0.23	0.36	0.62
H2OSOURC	5	0.11	0.27	09.0
FOXRABIT	5,8	0.32	0.42	0.68
BSOIL	3,5	0.20	0.32	0.62
CHEMCHNG	5	0.33	0.42	0.61
BPLLUT	3	0.31	0.41	0.63
BPOLAR	3	0.18	0.29	0.58
BSTORM	3	0.23	0.33	0.63
CONTOURX	8	0.29	0.38	0.58
CAVITY	8	0.42	0.49	0.63
YHUMID	3	0.23	0.33	0.54
BPLNT3	3	0.12	0.21	0.52
LAKE	8	0.26	0.34	0.57
TEMPMIX	8	0.15	0.23	0.50
PHYSPROP	5	0.27	0.34	0.57
NERVOUS	5	0.07	0.16	0.43
See notes at end of table.				

Science assessment estimated proportion correct: School years 2001–02, 2003–04, and 2006–07—Continued Table C3.

	Used in Grades	Round 5	Round 6	Round 7
BMAMML	3,5	0.11	0.20	0.43
PROPERTY	8	0.33	0.39	0.58
CREEK	8	0.31	0.36	0.55
SUNMOVE	Ś	0.23	0.29	0.48
CONSTELL	Ś	0.28	0.32	0.51
SOLUTION	5,8	0.20	0.26	0.44
PREDATOR	8	0.26	0.30	0.49
ALGAE	8	0.20	0.24	0.44
TEMPLOW	S	0.08	0.15	0.36
CELLS	8	0.41	0.46	0.59
PENCLH20	5,8	0.18	0.24	0.43
REPRODUC	8	0.18	0.23	0.41
PRECIPIT	8	0.26	0.30	0.46
STORM	8	0.17	0.22	0.39
BEARCUB	5	0.33	0.37	0.49
WHYFAST	5	0.08	0.13	0.27
GRANITE	8	0.09	0.11	0.23
H2ORECYC	5	0.38	0.42	0.50
LENS	∞	0.18	0.20	0.30
GUINEA	8	0.09	0.12	0.22

Science assessment estimated proportion correct: School years 2001–02, 2003–04, and 2006–07—Continued Table C3.

NOTE: IRT-estimated proportion correct for each item in each round. Estimates for third through eighth grade have been put on a common scale to support comparisons. Science was not tested in kindergarten/first grade. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). The grades in which items appeared on assessment forms are noted. Not all items appeared in test forms for all rounds. Table estimates are based on cross sectional-weights within each round (C5CW0, C6CW0, C7CW0). SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002, spring 2004, and spring 2007.

This page is intentionally left blank.

APPENDIX D

ECLS-K DIFFERENCE BETWEEN ACTUAL AND ESTIMATED PERCENT CORRECT BY ROUND

8-99, 1999-2000, 2001-02,	
l years 199	
and: School	
orrect by rou	
d percent co	
nd estimate	
en actual a	
ence betwe	
g assessment differ.	14, and 2006–07
Reading	2003-0
Table D1.	

	IRT "a"							
	parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
CANDLE	0.79	0.00	-0.01	0.00	-0.02	•;	•;•	
POURING	0.87	0.01	-0.01	-0.03	-0.06	•		•;
CEREAL	1.15	0.01	0.00	0.00	-0.04	•;	-;	
DECORATD	0.77	0.01	0.00	-0.01	-0.04	-;	-;-	
BEGBIKE	1.65	0.01	0.00	0.01	-0.01		-;	- -
BEGIN	06.0	0.01	0.00	0.01	-0.01	-;	-;-	
LETRECD	2.71	0.01	0.00	-0.01	-0.01	•;=	-;	•;
VEGETBLE	0.73	0.02	-0.01	0.02	-0.08	•;	•;	
LETRECF	3.08	0.01	0.00	-0.01	-0.01	•	•	•;
LETRECM	2.71	0.01	0.00	-0.01	0.00	•	•	•}
LETRECT	2.89	0.00	0.00	0.00	0.00	•;		- ;-
COULDNOT	0.90	0.02	-0.01	0.02	-0.06	•;	•;	•;
KAYLAFLY	0.66	0.01	0.00	0.01	-0.06	- ;-	-;	÷
NEXTLINE	1.11	0.00	0.00	0.04	-0.01	-;	-;	•;
STORYEND	1.28	0.02	0.00	0.00	-0.02	- ;-	-;	÷
BEGP	1.74	0.01	0.02	0.00	-0.02	•	•	÷
TIME	1.07	0.01	-0.01	0.02	-0.01	•;	•;	•
BEGR	2.33	-0.01	0.02	0.02	-0.01	•;	•;	•
BEGL	2.30	-0.01	0.02	0.02	-0.01	•	•	+
TRUNK	0.82	0.02	0.01	0.00	-0.09	•	•	-
AWARDING	1.00	0.01	0.01	0.05	-0.02	•	•	-
JOGGING	1.23	0.04	-0.01	0.03	-0.09	•;	•;	÷
See notes at end of table.								

~	= = 541							
	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
COULD	0.60	0.02	-0.01	0.01	-0.03	•¦=	•;	•}
ENDL	2.16	0.00	0.01	0.00	-0.01	•;=		
MOM	2.36	0.00	0.01	-0.01	0.02	•;=		•;•
ENDF	1.80	0.00	0.01	0.01	-0.01	•;=		•;•
YELLOW	1.91	-0.04	0.02	0.09	0.10	•;=		•;•
BEGB	1.42	0.01	0.01	0.01	-0.02		- ;-	•;
BEGWORD	0.86	0.06	0.02	0.03	-0.04		•;	•;=
ENDP	1.63	0.01	0.00	0.01	-0.01		- ;-	•;
QMARK	1.10	-0.08	0.01	0.06	-0.01	•;=		•;•
ENDD	1.67	0.01	0.00	0.01	0.00		- ;-	•;
YOU	2.72	-0.03	0.01	0.08	0.15	•;=		•;•
ORPIG	2.10	0.05	0.01	-0.01	-0.11	•;=		•;•
ORSAT	2.75	0.00	0.02	0.05	-0.11		- ;-	*-
ORTAIL	3.11	0.06	0.02	-0.01	-0.10	•;=		•;•
RUNS	3.32	-0.01	0.00	0.00	0.03	-0.01	•;	*
ORHAND	3.17	0.05	0.03	0.01	-0.13	+	•;	•;=
NEEDHOME	4.00	-0.01	-0.02	-0.01	0.12		•	*
WENT	3.21	0.00	0.00	-0.01	0.02	-0.01	•;	•;=
DOWN	3.50	0.00	-0.03	-0.01	0.05	-0.01	•	•;
BOYBIKE	3.50	-0.02	-0.02	0.01	0.02	+	•}	•;
JEEP	3.02	0.01	0.00	-0.01	0.02	-0.06	•}	•;
FISHING	4.00	-0.01	-0.03	-0.04	0.05	+	•}	+
CANINBAG	2.12	0.00	0.00	0.04	0.01	+	•}	•;
KITNBED	3.16	0.00	0.00	0.01	0.01	•;	•	•;
CATCH	3.50	0.01	-0.01	-0.03	0.02	+	•}	•;
MAKE	1.34	• •	•	0.04	-0.01	•	•;	•;•
KNOW	2.63	-0.03	-0.07	-0.05	0.02		•;	*
LIGHT	4.00	0.03	-0.04	-0.05	0.01	+	•}	•;
ELEPHANT	3.80	0.04	0.01	0.00	-0.01	**	÷	÷
See notes at end of table.								

Table D1. Reading assessment difference between actual and estimated percent correct by round: School years 1998–99, 1999–2000, 2001–02,

	IRT "a"							
	parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
KIMCAT	4.00	0.01	0.02	0.01	0.02	•;•	•] •	•;
GIRLRED	3.02	0.01	-0.01	0.01	0.01	•;=	- ;-	•;
BACKPACK	3.10	0.03	0.02	0.02	-0.01	0.00	0.02	•;
FLATTIRE	3.55	-0.04	-0.03	-0.04	0.02	•;=	- ;-	•;
LIKEDRY	4.00	0.05	0.00	-0.03	0.00	•;=	-;	•;•
WRONG	3.65	0.05	0.05	-0.03	-0.01	•;		*
LISTEN	3.96	0.04	0.01	0.01	0.00	0.00	-0.02	•;
RIDEBIKE	3.75	0.07	0.03	0.00	-0.01	0.00	0.01	•;•
CHOCCAKE	4.00	0.05	0.00	-0.02	0.00	•;•	•;	•;•
SIZES	4.56	0.03	0.03	0.01	0.00	0.00	-0.02	•;•
QUIET	3.57	• ;	•;=	0.01	0.00	-0.01	•;•	•;
DOGHOUSE	2.80	0.02	-0.01	0.01	0.01	•;•		-;
ENVELOPE	3.74	0.08	0.04	0.02	-0.02	-;	• ; •	•;•
COMPARSZ	2.69	- ;-	• ;-	•;=	•;	0.00		•;
KINDLETR	3.33	•;=	•;•	•;=	•;	0.00	0.01	*
THROUGH	2.84	0.06	0.00	0.01	0.02	0.00	-0.03	•}=
ADULT SZ	2.15	• -	•;	•}=	•;=	0.00		•}=
GROWUP	2.53	•;=	•;•	•;=	•;	0.02	-0.01	*
WHENPAST	1.97	÷	•;	•}=	•;=	0.00	0.00	•}=
WHENTOOK	3.78	÷	•;	•}=	•;=	0.00	0.03	•}=
GAVEWHAT	3.43	•;=	•;•	•;=	•;	-0.01	0.00	*
KNIGHT	1.59	÷	•;	•}=	•;=	0.00	0.04	•}=
DANGER	1.80	• -	•;=	•}=	•;=	0.00	÷	•}=
RAGE	3.55	0.05	0.01	-0.01	-0.01	0.03	÷	•}=
MARCHED	4.00	0.05	0.02	-0.02	0.00	•;	•}=	•;
CATNAME	2.65	+	•;	0.03	0.00	•;	+	•;
AUTHFEEL	4.38	- 	•;	+	•;•	0.00	0.00	•;
WTLESS	5.82	+	•;	0.00	-0.02	0.01	0.01	•;
SAMEHANG	2.88	+	*	+	÷	-0.01	0.00	•;
See notes at end of table.								

	10T "."							
	parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
TOIL	2.53	0.06	0.02	0.00	-0.01	0.04	•}	-}
KOALA IS	2.90	•;	•;	- 	•;•	-0.01		•;•
CORNER	2.65	0.03	0.00	-0.01	-0.01	0.04		•;
TVSHOW	3.83	•;=	•;=	- :	- ;-	0.02	-0.01	•;
OWNRNAME	2.64			0.00	0.00	•;=		•;
REQUIRE	4.00	•;=		0.01	-0.02	0.02		•;
TANZANIA	3.62	•;=		- :	- ;-	0.02	-0.02	•;
CAPTURE	2.97	0.04	0.00	-0.02	0.00	0.01	-;	•;=
KIND OFC	2.99	•;	•;	- 	•;•	0.00	-0.01	-;
WEB	2.06	0.04	-0.01	-0.01	0.00	0.02	-;	•;=
FACT	3.68	•;=	•;=	- :	- ;-	-0.01		•;•
BABYSIT	4.09		-;		•;•	-0.01	-;	•;•
MOISTURE	4.03	•;=		-0.01	-0.04	-0.02	0.02	•;•
ROBBER	3.43			- :- -	•;	0.00	-0.01	•;
NONFICT	2.14	+			÷	0.00	+ -	+
UNUSUAL	4.00	•	•	0.05	0.00	•}=	•	•;
MOTHER	1.55		- -	- 	÷		0.00	*
WORDARTH	2.81	•	•	•;	•;=	0.00	-0.04	•;
STRAGGLE	2.02		- -	•;	÷	0.00	•	*
TRUEBAB	2.87	•	•	•;	•;=	-0.01	•	•;
RECIPE	3.90	0.08	0.05	0.02	-0.03	•;	•	•;
THOSEDAY	2.14	•;	•;•	+	+	-0.02	0.08	•;
MAINPROB	1.79	•;	•;•	+	+	+	0.00	•;
PREDICT	3.86	•	•	•;	•;=	•}=	-0.01	•;
WHYROUND	2.82	•	•	•;	•;=	0.00	-0.03	•;
JAMMED	2.23	- ;-	- -		÷	0.01	-0.02	*
PURPOSE	1.69				•;	•;=		0.00
LINECLUE	3.33	•	•	•;	•;=	0.00	-0.01	•;
EXAMPLE	3.45	*	+	+	+	+	0.00	+-
See notes at end of table.								

	Round 7	+ -	0.02	•;•	•;•	•;•	•;	•;•		•;•		-0.01	•;•-	•;•	-;		0.00	0.03	0.00	0.01	•;		0.00	•;	0.01	•;•-	•;	-;	•;	0.05	
	Round 6	•;•	-0.06	-0.01	0.02	-0.01	-;	- ;-	+	0.00	0.00	•;•	-;	0.01	0.03	0.02	• ; •	-0.03	• ; •	-0.01	•;	0.00	- ;-	- ;-	0.00	0.01	- ;-	0.00	- ;-	-0.02	
	Round 5	•;		0.01	-0.01	0.01	0.00					•;	•;	•;	-0.12	-0.01	- ;-		•;		•;•	-0.02	•;		•;=	-0.02	0.00	-;	0.00	0.00	
	Round 4	-0.02	•;•	•;=	•;=	•;=	•;=	0.00	0.00	•;=	•;•		0.00	-0.04	•;•	•;•	•;•	•;•		•}•-	0.00	•;	•;=	0.03	• ;-	•;=	•;=	•;=	•;=		
	Round 3	0.01	-;	- ;-	- ;-	- ;-	•;•	0.01	-0.02	•;=		•;•	0.00	0.02	-;				•;•	-;	0.05	-;	- ;-	0.05	•;=		- ;-		- ;-	•	
	Round 2	0.02		•;=	•;=	•;=	•;•	•;=		•;=		-;	•;•	•;•	-;				•;•	-;	•;•	-;	- ;-	- ;-	•;=		- ;-		- ;-		
	Round 1	0.00	- ;-				- ;- -								-;					-;	•;•	-;			•;•	-;		-;			
IRT "a"	parameter	5.14	1.87	3.50	3.09	4.42	2.31	3.95	1.95	3.49	2.17	3.71	2.61	3.55	3.34	3.02	1.70	3.18	3.16	3.97	4.05	2.55	1.46	2.34	2.64	2.01	3.55	3.07	2.84	1.62	
		INGREDNT	OVATIONS	BOW	SURPRISE	TRAIN	FRICTION	MYSTERLY	MOREINFO	DR ROSE	IMP UNDR	DECLINE	APPROX	WAGES	TEARING	FEELSAFE	NEWPLANS	DEPART	THEME	SPRING	VICIOUS	3THINGS	TRIALS	MNIDEA-S	4CORNERS	DEHYDRAT	POINT	SLUDGE	TAKECARE	HOWFEEL	See notes at end of table.

	IRT "a"							
	parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
ONAHW	2.72	•;	-;	-0.05	0.05	•;•-	-;	-;
DOMESTIC	2.93	•;•	-;	-;	-;	-0.01	0.02	-}
LIKECHDR	2.98	•;=	-;		-;	-0.01	0.04	•;
APOSTRPH	2.12	•;=	-;		-;	0.00		•;
DIFFRNT	3.18	•;	•;•	•;	•;•	•;	0.00	•;
SLOW LRN	3.07	•]•	-;	•;	-;		0.00	•}
INFLUENT	2.75	-;	•;•-	- ;-	•;	-0.04	0.01	-;
DIFFROOM	1.85	•]•	•}	•;	-;	0.03	-0.04	•}
CRITCISM	3.82	•]•	•}	0.00	-0.11	-0.03	0.05	•}
WHY LEFT	3.63	•;=	•;	- :- -	• ; •	- ;-	-0.02	0.02
COMPASS	4.54		•;=-	•;•-	•;	- ;-	0.00	•;•
ABOUT	3.12	-;	•;•	- ;-	•;	-;	0.00	-;
SIM PROB	1.93	•]•	•;•	•;	-;	0.04	-0.03	-0.03
DIFFMONK	1.97	•]•	•}	•;	-;	0.00	-;	•;•
WHYNOT	3.14	•;•	•}•	•;	- ;-	0.04	-0.04	-0.01
DISLIKE	3.26	•;	•;•	•;	•;•	•;	•;•	0.00
STRANDS	1.85	0.03	0.01	-0.01	0.00	-0.01		•;•
SURVIVED	3.16	•;•	-;	-;	-;	-;	-;	0.01
PREFRNCE	2.09	•;•	-;	0.11	0.08	0.00	-0.02	•;
PROBSOLV	1.91	•;		•;	•;•-	0.07	-0.03	•;
DESCPURP	2.75	•;=			-;	- ;-	- :- -	0.00
MAJTHEME	2.89		- ;- -	- :-	•;	0.04	-0.02	•}
PAWNS	1.53		- ;-	- :-	•;	•;	+ -	0.00
RACHEL	2.71		- ;- -	- :-	•;	0.06	-0.02	•}
WHYCONTR	2.69		•;=		÷	- ;-	0.00	•;=
MICROWAV	2.97		- ;-	- :-	•;	-0.02	0.01	•;
AMBITIOU	3.00	- -	•;=	0.03	0.06	0.00	•;	•;=
SUPPORT1	1.79		•;=		÷	- ;-	-0.06	0.02
ALIGNMNT	2.72	*-	+	-0.04	-0.13	+	0.01	+
See notes at end of table.								

	Round 7	0.01	-0.02	-0.01	•}	•;	•;	0.01	•}•	•;		•;	•;	0.01	•;	0.00	0.00	0.00	•;	0.00	•;	0.00	•;		•;	0.02	•;	0.04	0.04	•;	
	Round 6	•;•	-0.02	0.03	-0.01	•;	0.00	-0.02	-0.01	0.00	0.00	0.00	0.00	- ;-	0.00	- ;-	-;		0.03	- ;-	0.00	0.01	0.03	0.00	0.00	-0.01	-0.01	-0.12	-0.03	0.02	
	Round 5	-;	0.02	•;=	0.03	•;	•;	0.01	0.00	•;=		•;	•;•				•;•		-0.03		•;•		-0.03		0.01	•;=	0.01	•;•	•;	-0.01	
	Round 4	- ;- -	-;	- ;-	•;	0.01	- ;-	- ;-		- ;-	-;	- ;-	•;	-;	-;	-;		-;	-;	-;		-;		-;	- ;-	•;	•;	•;	•;	+	
	Round 3	- ;- -	-;	+ -	•;	-0.02	+ -	+ -		+ -	-;	+ -	•;	-;	-;	-;	•;•	-;	-;	-;	•;•	-;	•;•	-;	+ -	•;	•;			+	
	Round 2	-;		•;	•;	•;	•;	- ;-		•;		•;	•;•				•;•				•;•		•;•		- ;-	•;=	•;	•;	•;	+	
	Round 1	-;		•;			•;	•;		•;		•;	•;•-												•;	•;=	•;			+	
IRT "a"	parameter	2.73	2.78	2.68	2.80	3.17	3.20	2.32	2.40	1.68	1.69	2.75	1.91	1.14	1.81	1.97	2.30	2.13	4.01	2.23	1.44	6.13	4.16	2.47	1.79	5.64	2.50	3.24	1.37	4.66	
		EXAMS	HELPPRB	ON MESA	HOWAUTH	DESCRPIG	MTPCOMP	HELPUND	COMPARWH	SUMMARY	LIKE/DIS	ERUPT2	AUTHOR	SHAREBOT	PSYCHLG	TASKS	EFFORTS	PROTECT	HOAX	FEATURE	ADVANCES	DOUBT1	MNIDEA-A	INSUFFIC	GUESS	DOUBT2	TRUECROP	MAINPURP	THEORY2	3DISMISS	See notes at end of table.

	Round 7	0.00	•;•	-0.01	0.01	0.00	- -	0.00	0.02	0.01	0.01	- -	0.01	0.01	0.01		0.00	
	Round 6	- ;- -	-0.01	0.05	- ;-	•;	0.00			•;			•;				+	
	Round 5	- ;- -	0.04	- ;-		- ;-	- ;-	- ;-	- ;-	- ;-	- ;-	0.00	•	- ;-	- ;-	0.00	÷	
	Round 4	•;•-	•;•	•;•	•;•	•;=	- ;-	- ;-	- ;-	•;=	- ;-	- ;-	- ;-	- ;-	- ;-	- ;-		
	Round 3	•;•-	•;	•;	•;	•;=	- ;- -	- ;- -	- ;- -	•;=	- ;- -	- ;- -	- ;-	- ;- -	- ;- -	- ;- -		
	Round 2	-;				•;•				•;•			•;=					
	Round 1	-;				•;•	•;	•;	•;	•;•	•;	•;	•;	•;	•;	•;	+	
IRT "a"	parameter	1.64	3.28	3.02	2.87	1.23	0.82	2.00	1.45	2.38	2.54	3.95	1.93	2.04	3.67	1.16	1.09	
		FIONA	BESTWAGM	TONE	RESULT	ALTRUIST	BELLGRNT	TRIFLES	ROBOTS	OVERLOOK	CATTLE	ANOMALY	HISTORIAN	SUPPORT2	CAUGHT	EMBOLISM	RVLTIONS	† Not applicable.

Positive numbers indicate a higher proportion of actual correct answers than were predicted by the IRT model; negative numbers correspond to actual proportions that were lower than estimates. Statistics illustrate IRT model fit, not population estimates, and are unweighted. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, spring 2004, and spring 2007. NOTE: Difference between actual percent correct for test takers who answered each item, and IRT-estimated percent correct for the same children. Not all items appeared in test forms for all rounds.
07, 200	3-04, and 2000-01							
	IRT "a"		- - -	- ډ	-	-	-	r F
	parameter	Round I	Round 2	Round 3	Round 4	Round 5	Round 6	Kound 7
2CRAYONS	1.91	0.00	0.00	-0.01	-0.01	•;	•; •	•;
3BANANAS	0.96	0.02	-0.01	-0.01	-0.07	•;=	•;	•;
SQUARE	1.13	0.02	-0.02	-0.04	-0.05	•;=	•;	•;=
NUMBER 4	3.50	0.00	0.00	-0.01	-0.01	•;=-	- ;-	•;
# STRAW	1.31	-0.02	0.00	0.01	0.00			•;
STICKBAT	1.09	0.03	-0.01	0.00	-0.03	•;•		-;
3-1PENCL	0.99	0.01	0.00	-0.01	-0.06	-;	-;	•;•
NUMBER 7	3.34	0.00	0.00	-0.02	-0.02	•;•		•;•
#VANILLA	1.47	0.00	0.00	0.02	-0.01	•;•		•;•
#CHOC	1.55	0.00	0.00	0.01	-0.01	•;•	-;	-;
NUMBER 9	2.86	-0.01	0.01	0.00	0.00	•;•	-;	-;
PNTBRUSH	1.85	0.00	0.01	0.01	0.00	-;	-;	•;•
COUNT 20	1.39	-0.01	0.03	0.00	-0.03			•;
4LINES	0.70	0.03	0.02	0.04	-0.06			•;
6BANANAS	1.35	0.00	0.01	0.05	0.03			+
LG-SM-SM	1.80	0.00	0.01	0.02	-0.01		•;•	•
SM-LG-SM	1.61	0.00	0.01	0.03	-0.01	+	•	•;=
NUMBER17	2.32	-0.02	0.03	-0.01	0.01	•;=	•;	•;=
X000	1.29	0.00	0.03	0.06	-0.01	+	•	•;=
NUMBER23	2.32	-0.01	0.01	-0.01	0.01	+	•	•;=
3RD LINE	2.23	0.01	0.01	0.01	-0.02	+	•	•;=
3+2 CARS	1.56	0.03	-0.01	0.00	-0.02	+	•	•;=
HALFOVAL	1.10	0.00	0.03	0.03	0.01	•;=	•;	•;=
78910	2.17	-0.04	0.02	0.04	0.01	•;=-	- ;-	•;
2+3STICK	1.73	0.01	-0.01	0.01	0.00		•;•	•;=
#BUGS	1.69	0.03	0.00	0.00	-0.02	•;	-;	•;=
2 + 2	3.00	-0.07	-0.02	0.02	0.04		•;•	•;=
3 + 3	4.00	-0.05	-0.02	0.01	0.02		•;•	•;=
1 + 7	1.61	0.00	-0.03	0.02	0.04	*	*	•;
See notes at end of table.								

Table D2. Mathematics assessment difference between actual and estimated percent correct by round: School years 1998–99, 1999–2000, 2001–

	Round 7	•;•	•;	•}=	•;	•;	•;	•;	•;	•;		•;•	•;•	•;	•	*	•;	•;	•;	*	•;	*	•}=	•}=	•}=	•;	•}=	•}=	*	*	
	Round 6	•;•	•;	•;=	•;	•;	•;	•;	•;	•;	-;	-;	-;	•;	•;	•;	•;	•;	•;	•;	•;	•;	•;=	•;=	-0.01	•;=	•;=	•;=	•;	-0.02	
	Round 5	0.01	0.00	•;•	- :- -		-0.16	- ;-	- ;-		0.00	-0.01	0.07	0.00	•	- ;-	- ;-	0.01	- ;-	- ;-	- ;-	0.00	•;•	0.01	0.00	0.00	-0.04	0.01	- ;-	0.01	
	Round 4	•;•-	•;	0.02	0.00	-0.02	0.01	-0.01	-0.01	-0.03	0.01	0.01	0.04	•;	0.01	0.04	0.04	•;	0.00	0.01	-0.04	•;=	0.00	•}=	0.04	•}=	0.01	•}=	0.02	-0.03	
	Round 3	-;	•;	-0.02	0.01	0.00	0.04	0.04	0.02	0.01	-0.03	-0.01	-0.04	-;	0.01	-0.09	-0.09	-;	-0.01	0.01	0.05	•;•	-0.02	•;	-0.04	•	0.02	•;	-0.02	0.04	
	Round 2	-;	- ;-	-0.01	0.00	0.00	0.01	0.01	0.00	0.01	-0.01	-0.01	-0.01		-0.01	-0.08	-0.08	-;	-0.03	-0.01	0.04	•;•	-0.02	•;	-0.03	•	0.00	•;	-0.01	0.02	
	Round 1	•;•	- ;-	0.00	-0.06	0.03	0.01	0.00	-0.01	0.02	0.00	0.02	-0.02	•;•	-0.02	-0.09	-0.09	•;•	-0.03	-0.01	0.09	•;=	0.03	+	-0.02	•}=	-0.01	+	-0.01	0.04	
IRT "a"	parameter	1.23	2.62	1.47	2.47	2.15	1.55	0.78	1.58	1.83	2.26	2.25	2.53	1.90	2.49	3.06	3.21	1.26	2.52	2.07	2.14	2.88	1.96	1.40	1.95	1.72	2.45	2.53	2.84	2.46	
		TEAMS_R	VICKS_R	8-6CRAYN	3 + 4	5-10RANG	2+5MARBL	SHAPES	PATTERN	2+5CIRCL	12 BY 2S	3+7PENNY	51015 25	ORANGE R	11 + 3	7 - 3	9 - 2	PATHS_R	6+7	12 + 6	# MORE	MOST_Y	2-1+2	RULER_R	$A13_79$	SIDES_R	4+4-2	PAGES_R	17 - 4	$COST_{10}$	See notes at end of table.

	IRT "a"							
	parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
12-9	2.79	-0.04	-0.05	-0.05	0.03	- !- -	•;•	•¦=
26 + 20	2.90	-0.01	-0.03	-0.04	0.02		•;=	•;
CARS15_5	2.58	0.03	0.02	0.02	-0.01	0.00	-0.02	+
FEWESTY	2.72		•;=	•;=	•;=	-0.01	- ;-	•
SQUARE R	1.04		•;=-	÷	•;	0.00		•;
CUBES10	1.01		•;	•;=	•;=	0.00	0.00	•
HOWMANY\$	1.73	0.04	0.01	0.02	-0.04	0.08		•;
CANDY8_2	2.53	0.02	0.02	0.04	-0.03	0.01	0.01	•;
BEADS_R	4.38	•;=	•;=	•;	•;	0.02	•;=	•;
12-? PEN	2.77	0.06	0.04	0.02	-0.03	0.03	•;=	•;
NEXT78	2.74	•;	•;•-	•;•	•;•	-0.01	0.01	-;
HEADSUP	1.33	0.05	0.00	0.00	-0.03	0.07	•;=	•;
24-14BKS	3.01	0.01	0.01	0.01	-0.01	•;		•;
MEANS_R	3.06	+	•;	÷	•;=	0.00	•	•
EQUAL_R	3.23	•;	•;	•}=	•;=	0.01	•;	•
DO_ADD4	2.44	+	•;=	- 	+	0.02	-0.03	•
MONEY_R	3.72	•;	•;	•}=	•;=	0.01	•;	•
TIME1030	2.19	+	•;	• -	•;=	0.00	0.01	•
POINTS_R	2.01			•;=	•;=	0.01	- ;-	•
SCORE_Y	3.38	+	•;	• -	•;=	-0.01	•	•
GOALS	2.47	0.00	0.00	0.00	0.00	0.00	•;	•
PAPER	3.04	+	•}=	•	•}	0.00	•	•
NICKELS	2.63	+	•;	• -	•;=	0.00	•	•
17CENTS	3.08	0.00	0.00	-0.01	0.01	•	•	•
MORE1_Y	3.90	•;	•;	•}=	•;=	0.00	•;	•
NUMBER60	3.66	+	•;	- 	+	0.00	0.00	•
BDCAKE	2.39	-0.01	-0.01	-0.01	0.01	•;•	•	•
CUBESIDE	1.45	+	•;	- 	+	-0.01	0.01	•
NEXT120	3.12	•;	*	+	+	0.00	0.00	•;
See notes at end of table.								

	Round 7	•;•	0.00	•;=	•	•;=	•;=	-0.01	•;•-	-;	•;•	•}	•}	•}	-0.02	•;•-	•}	0.00	•;•-	•;•	-;	0.02	•}	•;•	•;	0.06	0.01	0.00	0.00	+	
	Round 6	•; •	-;	-0.01	0.00	0.00	- -	0.00	0.00	0.00	0.03	- ;-	- ;-	- ;-	-0.01	0.00	- ;-	- ;-	0.02	0.02	-0.01	-0.01	- ;-	-;	0.00	0.03	- -	÷	÷	0.02	
	Round 5	0.00		0.00	•;	0.00	0.00	0.01			-0.06	0.01	-0.01	0.00	0.02		-0.01		-0.01				0.00		•;•	-0.05	•;	•;	•;	-0.04	
	Round 4	•;•	•;	•;=	•;=	•;=	•;=	•;	•;	-;	•;	•;	•;	•;	•;	•;	•;	•;	•;	•;	-;	•}•-	•;	-0.01	•;	•;	•;=	•;=	•;=	+	
	Round 3	-;		•;•	•;=	•;•	•;•	•;=		-;		•;=	•;=	•;=			•;=	•;=			-;		•;=	0.01	•;•		•;•	•;•	•;•	+	
	Round 2	•; •	-;	- ;-	•;	- ;-		- -	- ;-		- ;-		•;	•;	- ;-	-;	•;	•;	- ;-	- ;-	-;			0.02	•;	÷			- ;-	+	
	Round 1	•;•-	-;		•;			- ;-	-;	•;•-	-;	- ;-	- ;-	- ;-	-;	-;	- ;-	- ;-	-;	-;	•;•-	•;•	- ;-	0.02	•;	÷				+	
IRT "a"	parameter	3.91	2.93	2.15	1.82	3.67	2.69	2.66	3.21	3.02	2.18	3.01	2.82	1.43	2.79	1.94	2.65	4.01	2.89	2.19	2.38	1.61	2.05	1.74	2.08	2.27	1.86	1.32	1.74	2.45	
		FEWER Y	TREES -MC	CHART_64	AGEGRAPH	BOX_700	NUMBER	TBSP7OZ	CANDY27	TREES100	COLORSYM	FRIES	CHILDR Y	STAR-Y	PGSLEFT	BOXSHELF	SECOND	SAMJUAN	A568214K	1ST#X5	BIKETIME	MISSNUM	FRUIT	24/4 TAB	SCALE=	CARWASH	SHANBIKE	FLIPTRI	SCALEMAR	CARDS579	See notes at end of table.

	IRT "a"							
à	arameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
	2.29	•;•	-;	•;•	•;	0.05	0.01	-0.04
	2.59	- -	-;	-;		•;•	0.00	•;
	1.62		•;	•;=		0.00	- 	*
	1.94		•;	•;=		•;=	- 	0.00
	2.47	•;=		•;•	•;=	-0.03	0.05	0.00
	1.87	•;=	•;	•;•	•;=	0.00	•;=	
	2.01		•;	•;=		0.00	- 	•;
	1.96	•;=	•;	•;•	•;=	- ;-	•;=	0.00
	1.53	•;	-;	-;		•;•		0.00
	2.09	•;•		•}•	•;	0.02	-0.02	-0.01
	1.29	•;	-;	-;		0.00	• !	•;•
	3.04	•;	-;	-;		0.02	-0.01	•;•
	1.96	•;=	•;	•;•	•;=	- ;-	-0.03	0.03
	2.57		•;	•;=		•;=-	- 	0.00
	2.13			•;	+	0.00	•;=	•;
	3.48	+	•;	•}=	•	•;=	0.00	•;
	4.16	+	•	•}=	•	•;=	0.00	•
	2.05	+	•;=	•}=	•	•;=	0.00	•
	2.12			•;=		•;=-	- 	0.01
	2.35	•;=		•;•	•;=	0.00	•;=	-;
	2.25		•;=	•;=			0.00	•;=
	2.02	+	•	•}=	•	0.04	-0.01	-0.03
	2.95	+	•	•}=	•	-0.06	0.03	•;
	1.91	•;	•;=	•}=	•;	+	•;	0.01
	1.81	•;		+	•;	+	•;	0.00
	2.00	0.04	0.00	0.00	0.00	•;	+ - -	•;
	2.97	•;	•	•}=	•;	-0.01	0.04	•}=
	2.00	•;	+	+	•;	-0.01	0.00	+
	2.65	*	- ;-	+	•;	*	*	0.00

	Round 7	•;•	0.03	•;	- ;-	0.01	-;	-0.04	0.00	0.03	-;	0.00	0.01	0.01	0.04	•; -	- ;-	0.00	+ -	•;•	•;	0.12	0.00	0.01	•;•	0.01	0.03
	Round 6	•;•	-0.02	0.00	0.00	-0.01	0.02	0.06	- ;-	-0.02	0.00	-;	-0.02	- ;-	-0.02	-0.01	0.01	0.01	÷	0.00	0.00	-0.14	•	•;•	0.00	•;	-0.02
	Round 5	0.00	-0.01	• ;-	•;	•;	-0.02	0.03	•}=	•;	•;	-;	•;	•;	•;	•;=	•;	•;	0.00	•;=	• ;-	•;=	•	•;=	•;=	•;=	•;
	Round 4	•;•	÷	•	÷	÷	÷	- ;-	÷	÷	÷	- -	- ;-	÷	- ;-	÷	÷	- ;-	÷	÷	•	÷	•}=	÷	÷	÷	•;
	Round 3	• ;-	•;=	÷	- ;-	- ;-	- ;-	•;=	- ;-	- ;-	- ;-	•;•-	•;=	- ;-	•;=		- ;-	•;=		+	÷	+	+	+	•;	•;	- -
	Round 2	•;•	•;=	- ;-	- ;-	- ;-	- ;-	•;	- :-	- ;-	- ;-	•;•-	•;	- :-	•;=		- :-	•;	•;=	• ;-		+	•;	+	•}=	•}=	- -
	Round 1	• ;-	-;	+ -				•;=				•;	•;=		•;=	-;		•;=		+	+ -	+	+	+	•;	•;	- -
IRT "a"	parameter	2.79	1.45	2.72	2.70	3.39	2.32	2.08	2.71	3.16	3.05	2.80	2.85	1.89	2.66	3.00	1.42	2.79	2.41	4.00	2.81	2.73	2.65	2.28	2.51	1.51	3.20
		MYSTER_B	NUMEDGES	OJ 300Z	FRAME3FT	SAND40MC	SAMEFRAC	MRKRULER	NUMSENT5	SMITHFAM	SHADED.2	X5PLUS3X	3FRAC4	LIFEEXP	RECPLAY	OPOSITIV	SALESTAX	PIZZA	FENCE_B	SHADED.3	DIFF=88	VOLPRISM	500BATT	CROSSLIN	MEASDIAM	NPLUS5	CARPSFOE

See notes at end of table.

	IRT "a"							
	parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
CIRCCIRC	1.98	•;	•;•	•] •	•;•	•;•	-;	0.00
12INCH	2.03		- ;-	- ;-	- ;-	0.00	0.00	0.00
FOLDCUBE	1.71		•;	•;			•;	0.00
† Not applicable.								

Positive numbers indicate a higher proportion of actual correct answers than were predicted by the IRT model; negative numbers correspond to actual proportions that were lower than estimates. Statistics illustrate IRT model fit, not population estimates, and are unweighted. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2000, spring 2007. NOTE: Difference between actual percent correct for test takers who answered each item, and IRT-estimated percent correct for the same children. Not all items appeared in test forms for all rounds.

	IRT "a"			
	parameter	Round 5	Round 6	Round 7
RBULB	0.91	00.0	• ;•	•;-
RENRGY	1.30	0.00	• !•	•;•
RPLANT	1.49	-0.01	-;	
RORGAN	0.58	0.01		•;
RTOOL	0.93	0.00	-;	
ROUIMM	1.04	0.00	0.00	•;•
RDSAST	1.51	0.00	•; -	•;
RFGRPS	0.62	0.01	-;	
RFORMS	0.97	0.00	•;•	•;•
YPLAIN	0.79	0.00	-;	
RWINGS	1.63	0.02	-0.08	•;
RANIML	0.96	0.00	-;	- :-
ROUFRZ	1.51	-0.01	0.00	-;
ROCCUR	1.59	0.01	•; -	•;
WHCHPREY	1.26	•;=	0.00	-;
RSEEDS	0.94	0.00	•; -	-;
ROUTAP	0.62	-0.02	0.02	- :-
ROUJUN	1.44	0.00	-0.01	•;•
RTHING	1.21	0.00	- ;-	•;•
RWATER	1.12	0.01	-;	•;•
YDSAST	0.83	0.00	-;	•;•
RSUNIS	1.23	0.01	•;•	+
SOLAR	0.92	-;	•;•	-0.01
ROUBRN	1.44	0.00	-0.01	+
ROUERT	0.88	0.02	-0.01	-0.13
RFISHB	1.06	0.00	•;•	+
RSHAPE	1.14	0.00	•;•	+
RHEART	1.35	0.00	0.01	•;•
RPWDER	1.03	0.01		÷
See notes at end of table.				

Science assessment difference between actual and estimated percent correct by round: School years 2001–02, 2003–04, and 2006–07 Table D3.

	IRT "a"			
	parameter	Round 5	Round 6	Round 7
ROUJAR	0.63	0.00	-0.01	
CUTSCAB	1.06	-;	0.00	0.00
ROUSRF	1.24	0.00	0.00	•;•-
RDESRT	1.05	0.01	-0.02	•;
PYRAMID	0.68	-;	0.05	-0.03
MONSNTM	1.08	-;	0.01	•;
YTHEMT	0.73	0.03	-0.01	-0.07
BEARTH	0.77	0.00	•; -	•;
SUGARDIS	0.99	-;	0.01	•;
YSOUND	0.99	0.00	•;•	•;
YINSCT	1.27	0.00		•;
YMOON	1.38	0.03	-0.02	-0.03
YSENSE	1.04	0.00	- ;-	•;
EARTHQK	1.48	•;•	0.02	-0.06
GRAVMOON	1.17		0.01	-0.01
PROTECT	1.30	•;	0.00	+
BSHADW	1.13	-0.01		+
ROUSOL	0.85	-0.04	0.05	•;
THUNDER	1.28		-0.01	0.04
AIRPOLL	1.51	•;	0.00	•;
WATRGRPH	0.88		-0.02	0.00
YBEES	1.13	0.01	-0.03	+
ROUBLB	0.00	-0.01	0.02	•;
ROUGRT	1.14	0.00	0.01	•;
ROUMTN	1.60	-0.02	0.03	-0.03
ROUMCE	1.51	0.01	-0.01	•;
ROUFLY	1.35	-0.01	0.01	•;
ECLIPSE8	0.66	•;		0.00
BSOUND	0.93	-0.07	0.02	0.01
See notes at end of table.				

D3. Science assessment difference between actual and estimated percent correct by round: School	vears 2001–02, 2003–04, and 2006–07–Continued
Table D3.	

parameter Rc YDSOLV 0.80 LAMPWIRE 0.96 ROUSHD 0.87 MIXTURE 0.87 MIXTURE 0.87 MIXTURE 0.87 MIXTURE 0.87 MIXTURE 0.87 MIXTURE 1.42 BPLNT2 1.15 YFWATE 1.41 BPLNT2 1.15 YLIVE 1.16 PLNT2 1.15 YLIVE 1.16 PRLANT 1.17 BPLANT 1.17 BPLANT 1.13 BPLANT 1.25 YFARMG 0.52 BPLANT 1.25 BPLANT 0.76 BURED 0.76 UNESTON 0.76 BURED 0.70 LIMESTON 0.76 BURED 0.71 BURED 0.70 BOLAR 0.71 BPLLUT 0.85 BPLL	Round 5		
YDSOLV 0.80 LAMPWIRE 0.96 ROUSHD 0.87 MIXTURE 0.42 ROUSHD 0.87 MIXTURE 1.42 BPLNT2 1.41 YFWATE 1.41 BPLNT2 1.15 YLIVE 1.16 YLIVE 1.04 YLIVE 1.17 BPLANT 1.25 YLIVE 1.17 BPLANT 1.25 YLIVE 0.76 UNTED 1.13 BPLANT 1.25 YFARMG 0.76 BHIBER 0.76 YFARMG 0.52 BHIBER 0.76 VFARMG 0.76 BURED 0.76 UNESTON 0.71 BURED 0.70 BURED 0.74 BURED 0.74 BURED 0.74 BURED 0.74 BSOIL 0.74 BPLLUT	000	Round 6	Round 7
LAMPWIRE 0.96 ROUSHD 0.87 MIXTURE 1.42 ROUSHD 0.87 MIXTURE 1.42 ECLIPSES 1.15 YFWATE 1.41 BPLNT2 1.15 YLIVE 1.04 YLIVE 1.04 YLIVE 1.04 YLIVE 1.04 PPLANT 1.25 YLIVE 1.17 BPLANT 1.13 BPLANT 1.13 BPLANC 0.76 PRLANC 0.76 BHIBER 0.76 YFARMG 0.76 BURIED 1.13 BURIED 0.76 BURIED 0.76 BURIED 0.70 BURIED 1.20 BURIED 1.21 BURIED 0.76 BURIED 0.74 BURIED 0.74 BURIED 1.20 BURIED 0.74 BOLAR 0.74 BPLUT 0.85 BPL	0.00	•;•	-;
ROUSHD 0.87 MIXTURE 1.42 ECLIPSE5 1.15 YFWATE 1.41 BPLNT2 1.15 YLIVE 1.17 BPLNT2 1.16 YLIVE 1.17 BPLANT 1.17 YLIVE 1.17 BPLANT 1.17 YLIVE 1.17 BPLANT 1.17 YLIVE 1.17 BPLANT 1.17 YBLANC 0.04 BHIBER 0.76 YFARMG 0.76 BHIBER 0.76 YFARMG 0.76 BURIED 0.76 BURIED 0.76 BURIED 0.71 BURIED 0.71 BURIED 0.74 SEEDGROW 0.71 BOLL 1.27 BSLIDE 0.74 BOLL 1.20 BPLLUT 0.85 BPLLUT 0.75 CONTOURX <td>-;</td> <td>0.00</td> <td>•;</td>	-;	0.00	•;
MIXTURE 1.42 ECLIPSE5 1.15 YFWATE 1.41 BPLNT2 1.04 YLIVE 1.07 DPLANT 1.25 YLIVE 1.17 BPLANT 1.25 YBLANC 1.13 BPLANT 1.25 NBLANC 0.90 BPLANT 1.25 NFARMG 0.76 0.76 0.70 LIMESTON 0.70 LIMESTON 0.70 LIMESTON 0.70 DOCONOURC 1.31 H20SOURC 1.31 H20SOURC 1.31 FOXRABIT 1.27 BSLIDE 0.70 DIAR 1.31 FOXRABIT 1.27 BSOLL 1.27 BSOLL 1.27 BSOLL 1.27 BSOLL 1.27 BSOLL 1.20 BPOLAR 1.20 BPOLAR 1.20 DIAR	0.04	-0.05	-;
ECLIPSE5 1.15 YFWATE 1.41 BPLNT2 1.04 YLIVE 1.50 YLIVE 1.50 YLIVE 1.17 BPLANT 1.17 YLIVE 1.17 BPLANT 1.17 YLIVE 1.17 SPLANT 1.17 BPLANT 1.17 BPLANT 1.17 BPLANT 1.17 BPLANT 1.17 BPLANC 0.76 UIMESR 0.76 VFARMG 0.52 BHIBER 0.76 VFARMG 0.52 BURIED 0.76 BURIED 0.70 LIMESTON 0.70 BURIED 0.71 BSLIDE 0.74 SEEDGROW 0.71 BSCIL 1.27 BSOIL 0.82 CHEMCHNG 0.85 BPOLAR 1.20 BPOLAR 0.75 CONTOURX 0.75 CAVITY 0.75 <td>+</td> <td>0.01</td> <td>-;</td>	+	0.01	-;
YFWATE 1.41 BPLNT2 1.04 YLIVE 1.17 VLIVE 1.50 CUPTEMP 1.17 BPLANT 1.25 YBLANC 1.17 BPLANT 1.25 VILIVE 1.17 BPLANT 1.25 VBLANC 0.76 BHIBER 0.76 YFARMG 0.52 EARTHCTR 0.76 BHIBER 0.76 VFARMG 0.52 EARTHCTR 0.76 BURIED 0.70 LIMESTON 0.70 BURIED 0.70 LIMESTON 0.70 BURIED 0.71 BURIED 0.71 BSLIDE 0.74 SEEDGROW 0.82 H2OSOURC 1.31 FOXRABIT 1.27 BSCILL 0.85 BPOLAR 0.66 BPOLAR 0.75 CONTOURX 0.75 CAVITY 0.75	+	0.00	
BPLNT2 1.04 YLIVE 1.17 YLIVE 1.17 BPLANT 1.17 BPLANT 1.17 BPLANT 1.13 BPLANT 1.13 BPLANT 1.13 BPLANT 1.13 BPLANT 1.13 BPLANC 0.76 BPLANC 0.76 BPLANC 0.76 BPLANC 0.76 VFARMG 0.52 EARTHCTR 0.90 BURIED 0.76 BURIED 0.70 LIMESTON 0.70 BURIED 0.710 BSLIDE 0.74 SEEDGROW 0.82 H2OSOURC 1.31 FOXRABIT 1.27 BSCILL 0.85 BSOIL 0.76 CHEMCHNG 0.66 BPOLAR 1.20 BPOLAR 1.20 BSTORM 0.75 CONTOURX 0.75 CAVITY 0.75 <td>0.00</td> <td>-;</td> <td>•;•</td>	0.00	-;	•;•
YLIVE 1.50 CUPTEMP 1.17 BPLANT 1.25 YBLANC 1.13 BHIBER 0.76 YFARMG 0.76 BURIED 0.76 BURIED 0.70 BURIED 0.70 BURIED 0.70 BURIED 0.70 BURIED 0.70 BURIED 0.74 BURIED 0.74 BURIED 0.71 BURIED 0.73 BURIED 0.74 BSLIDE 0.74 SEEDGROW 1.23 H2OSOURC 1.31 FOXRABIT 1.27 BSLIDE 0.82 BSLIDE 0.82 BSUL 1.27 BSOIL 1.27 BSOIL 1.27 BPOLAR 1.20 BPLUT 0.85 BPLUT 0.75 CONTOURX 0.75 CAVITY 0.75	-0.02	0.01	
CUPTEMP 1.17 BPLANT 1.25 YBLANC 1.13 BHIBER 0.76 YFARMG 0.52 EARTHCTR 0.52 BURIED 0.76 BURIED 0.70 BURIED 0.82 H2OSOURC 1.21 H2OSOURC 1.31 FOXRABIT 1.27 BSLIDE 0.82 BSCOIL 1.31 FOXRABIT 1.27 BSOIL 0.66 BPLLUT 0.85 BPLLUT 0.85 BPLLUT 0.85 BPOLAR 0.75 CONTOURX 0.75 CAVITY 0.75	0.01	•;=	-;
BPLANT 1.25 YBLANC 0.76 YFARMG 0.76 BHIBER 0.76 YFARMG 0.52 EARTHCTR 0.90 BURIED 0.70 LIMESTON 0.70 BURIED 0.70 LIMESTON 0.70 BURIED 0.70 LIMESTON 0.70 BSLIDE 0.74 SEEDGROW 0.82 H20SOURC 1.02 FOXRABIT 1.31 FOXRABIT 1.31 BSOIL 0.82 BSOIL 0.85 BPOLAR 0.66 BPOLAR 0.54 YHUMID 0.75 CAVITY 0.75	- 	0.01	
YBLANC 1.13 BHIBER 0.76 YFARMG 0.52 EARTHCTR 0.90 BURIED 0.52 EARTHCR 0.90 BURIED 0.70 LIMESTON 0.70 BURIED 0.70 LIMESTON 0.70 BURIED 0.70 LIMESTON 0.70 BURIED 0.74 SEEDGROW 0.82 H20SOURC 1.02 PSCIL 1.31 FOXRABIT 1.27 BSOIL 0.85 BSOLL 0.66 BPLLUT 1.20 BPLLUT 0.85 BPOLAR 1.20 BSTORM 0.75 CONTOURX 0.75 CAVITY 0.75	0.08	-0.02	
BHIBER 0.76 YFARMG 0.52 EARTHCTR 0.90 BURIED 0.70 BULDE 0.74 SEEDGROW 1.02 H2OSOURC 0.82 H2OSOURC 1.31 FOXRABIT 1.31 FOXRABIT 1.31 FOXRABIT 0.82 BSOIL 1.27 BSOIL 1.27 BPLUT 0.66 BPLUT 0.85 BPLUT 0.85 BPLUT 0.54 YHUMID 0.75	0.01	•;=	-;
YFARMG 0.52 EARTHCTR 0.90 BURIED 0.70 BURIED 0.70 BURED 0.70 BSLIDE 0.74 SEEDGROW 1.02 H2OSOURC 0.82 H2OSOURC 1.31 FOXRABIT 1.31 FOXRABIT 1.31 FOXRABIT 1.31 FOXRABIT 1.31 FOXRABIT 1.31 FOXRABIT 1.31 BSOIL 0.66 BPLLUT 0.85 BPLLUT 0.85 BPLLUT 0.85 BPOLAR 1.20 BSTORM 0.75 CONTOURX 0.75 CAVITY 0.75	0.00	-;	-;
EARTHCTR 0.90 BURIED 0.70 LIMESTON 0.70 LIMESTON 0.70 BSLIDE 0.74 BSLIDE 0.74 BSLIDE 0.74 BSLIDE 0.74 BSLIDE 0.74 SEEDGROW 1.02 H2OSOURC 0.82 H2OSOURC 1.31 FOXRABIT 1.31 FOXRABIT 1.31 FOXRABIT 1.27 BSOIL 1.27 BSOIL 0.66 BPLLUT 0.85 BPLUT 0.85 BPOLAR 1.20 BSTORM 0.75 CONTOURX 0.75 CAVITY 0.75 YHUMID 0.75	0.00	-;	-;
BURIED 0.70 LIMESTON 1.02 LIMESTON 1.02 BSLIDE 0.74 SEEDGROW 0.82 H2OSOURC 0.82 FOXRABIT 1.31 FOXRABIT 1.31 FOXRABIT 0.82 BSOIL 0.82 CHEMCHNG 0.66 BPLLUT 1.38 BPOLAR 1.20 BPOLAR 1.20 BSTORM 0.75 CONTOURX 0.75 CAVITY 0.75 YHUMID 0.75	+	-;	0.00
LIMESTON 1.02 BSLIDE 0.74 SEEDGROW 0.82 H2OSOURC 0.82 H2OSOURC 1.31 FOXRABIT 1.27 BSOIL 1.27 BSOIL 1.27 BSOIL 1.20 BPLLUT 0.85 BPLLUT 0.85 BPLLUT 0.85 BPLLUT 0.54 CAVITY 0.75 CAVITY 0.75 CAVITY 0.75	÷	0.01	-0.01
BSLIDE 0.74 SEEDGROW 0.82 H2OSOURC 1.31 FOXRABIT 1.31 FOXRABIT 1.31 FOXRABIT 1.31 FOXRABIT 0.82 BSOIL 1.31 CHEMCHNG 0.66 BPLLUT 0.85 BPLLUT 0.85 BPLLUT 0.85 BPLLUT 0.56 BPLUT 0.55 CONTOURX 0.75 CAVITY 0.75 YHUMID 0.75		•;	0.00
SEEDGROW 0.82 H2OSOURC 1.31 FOXRABIT 1.27 FOXRABIT 1.27 FOXRABIT 1.27 BSOIL 1.27 BSOIL 0.66 BPLLUT 0.85 BPLLUT 0.85 BPLLUT 0.85 BPLLUT 0.56 CHEMCHNG 0.66 BPLLUT 0.85 CHEMCHNG 0.66 BPLLUT 0.85 BPLLUT 0.85 CHEMCHNG 0.75 CONTOURX 0.75 CAVITY 0.75 YHUMID 0.75	-0.07	0.03	-0.01
H2OSOURC 1.31 FOXRABIT 1.27 BSOIL 1.38 CHEMCHNG 0.66 BPLLUT 0.85 BPLLUT 0.85 BPOLAR 1.20 BSTORM 1.81 CONTOURX 0.75 CAVITY 0.54 YHUMID 0.75	÷	0.00	0.00
FOXRABIT 1.27 BSOIL 1.38 CHEMCHNG 0.66 BPLLUT 0.85 BPLLUT 0.85 BPOLAR 1.20 BSTORM 1.81 CONTOURX 0.75 CAVITY 0.54 YHUMID 0.75		0.00	•
BSOIL 1.38 CHEMCHNG 0.66 BPLLUT 0.85 BPLLUT 0.85 BPOLAR 1.20 BPOLAR 1.20 BSTORM 1.81 CONTOURX 0.75 CAVITY 0.54 YHUMID 0.75		-0.10	0.06
CHEMCHNG 0.66 BPLLUT 0.85 BPOLAR 1.20 BPOLAR 1.20 BSTORM 1.81 CONTOURX 0.75 CAVITY 0.54 YHUMID 0.75	-0.04	0.02	-;
BPLLUT 0.85 BPOLAR 1.20 BSTORM 1.81 CONTOURX 0.75 CAVITY 0.54 YHUMID 0.75		0.00	-;
BPOLAR 1.20 BSTORM 1.81 CONTOURX 0.75 CAVITY 0.54 YHUMID 0.75	0.00	•;	-;
BSTORM 1.81 CONTOURX 0.75 CAVITY 0.54 YHUMID 0.75	0.01	•;	+
CONTOURX 0.75 CAVITY 0.54 YHUMID 0.75	0.01	•;	+
CAVITY 0.54 YHUMID 0.75	÷	-;	0.00
YHUMID 0.75		•;	0.01
	0.00	•;	•}=
BPLNT3 1.35	0.01	÷	*

	IRT "a"			
	parameter	Round 5	Round 6	Round 7
LAKE	1.04	•;•	•;•	0.00
TEMPMIX	1.25	•;•	-;	0.00
PHYSPROP	1.41	•;•	0.01	•;=
NERVOUS	1.03	•;•	0.01	•;•
BMAMML	0.78	0.05	-0.02	•;•
PROPERTY	1.05	•;=	-;	0.00
CREEK	1.16	•;=		0.00
SUNMOVE	0.99	•;=	0.01	•;=
CONSTELL	1.16	•;=	0.01	•;=-
SOLUTION	0.84	•;=	-0.01	0.01
PREDATOR	1.23	•;=		0.00
ALGAE	1.21	•;=	-;	0.00
TEMPLOW	0.78	•;=	0.01	•;=
CELLS	0.67	•;=		0.00
PENCLH20	0.88	•;=	0.00	0.00
REPRODUC	0.98	•;=	-;	0.00
PRECIPIT	0.94	•;=		0.00
STORM	0.91	•;=		0.00
BEARCUB	0.52	•;=	0.00	•;=
WHYFAST	0.57	•;=	0.00	•;=
GRANITE	1.24	•;=		0.01
H2ORECYC	0.43	• =	0.00	•;=
LENS	0.84		-}	0.01
GUINEA	0.68			0.00
4 NT 4				

Science assessment difference between actual and estimated percent correct by round: School years 2001-02, 2003-04, and 2006-07-Continued Table D3.

† Not applicable.

Not all items appeared in test forms for all rounds. Science was not tested in kindergarten/first grade. Positive numbers indicate a higher proportion of actual correct answers than were predicted by the IRT model; negative numbers correspond to actual proportions that were lower than estimates. Statistics illustrate IRT model fit, not population estimates, and are unweighted. Items are sorted in estimated ascending order of overall difficulty NOTE: Difference between actual percent correct for test takers who answered each item, and IRT-estimated percent correct for the same children.

(IRT "b" parameter). SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002, spring 2004, and spring 2007.

This page is intentionally left blank.