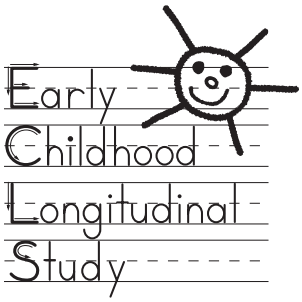
National Center for Education Statistics



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

Psychometric Report for the Third Grade

U.S. Department of Education
Institute of Education Sciences
NCES 2005–062

August 2005

Judith M. Pollack
Donald A. Rock
Michael J. Weiss
Educational Testing Service

Sally Atkins-Burnett
University of Toledo

Karen Tourangeau
Westat

Jerry West
Elvira Germino Hausken
National Center for Education Statistics

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Statistics
Grover J. Whitehurst
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

August 2005

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

This publication is only available online. To download, view, and print the report as a PDF file, go to the NCES World Wide Web Electronic Catalog address shown above.

Suggested Citation

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

Content Contact

Elvira Germino Hausken
(202) 502–7352
elvira.hausken@ed.gov

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

List of Appendixes

List of Tables

TABLE OF CONTENTS (continued)

List of Tables (continued)

TABLE OF CONTENTS (continued)

List of Tables (continued)

TABLE OF CONTENTS (continued)

List of Tables (continued)

TABLE OF CONTENTS (continued)

List of Tables (continued)

List of Tables (continued)

Appendix A Tables

TABLE OF CONTENTS (continued)

Appendix A Tables (continued)

TABLE OF CONTENTS (continued)

Appendix A Tables (continued)

TABLE OF CONTENTS (continued)

Appendix A Tables (continued)

Appendix B Tables

List of Exhibits

List of Figures

List of Figures (continued)

1. INTRODUCTION

This report documents the design, construction, and psychometric characteristics of the assessment instruments used in the spring 2002 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, 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.

This report documents the psychometric results for the fifth round of data collection, in spring 2002, when the majority of the sampled children were in third grade. A review of salient features of the kindergarten and first grade assessments is also included.

Two domains are represented by the ECLS-K third grade assessment instruments: cognitive (direct and indirect) and socioemotional. Direct cognitive measures refer to scores based on children's "direct" responses to cognitive test items. In third grade, direct cognitive tests were administered in reading, mathematics, and science. Indirect cognitive measures were ratings by teachers of the children's cognitive performance in the areas of language and literacy, mathematical thinking, science, and social studies. The socioemotional measures were teachers' ratings of children's social skills and approaches to learning. A questionnaire administered to the children included both indirect cognitive measures (self-ratings of competence in reading, mathematics, and all school subjects) and socioemotional questions relating to peer relationships and problem behaviors.

The direct cognitive assessments for third 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 reading and mathematics on vertical score scales based on successive assessments. (A general knowledge test had been given in kindergarten and first grade, but was replaced by a science assessment in third grade, so no vertical score scale is available for these subjects.)

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 knowledges that mark ascending critical points on the developmental growth curve. These multiple criterion-referenced levels serve two functions. First, they help with respect to the interpretation of 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 impact of certain school processes on changes in mastery of specific skills.

The development of the direct cognitive battery was carried out in five steps:

1. A background review was carried out of all the currently available psychometric instruments and the constructs that they purported to measure.

2. Test specifications were developed that were appropriate to the domains and constructs considered relevant for each grade.

3. Item pools were developed that reflected the test specifications in step 2.

4. The item pools were field tested in order to gather statistical and psychometric evidence as to the appropriateness of the items for carrying out the overall assessment goals.

5. The final test forms were assembled consistent with field test item statistics and the test specifications.

Chapter 2 of this report describes the objectives and design of the third grade assessment instruments. Differences between the kindergarten/first grade instruments and the third grade assessment battery 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 and first grade and the extension of these levels for third grade skills. Chapter 2 also describes the need for a small sample of second graders to bridge the gap between the spring-first grade and spring-third grade rounds for the purpose of establishing longitudinal score scales. For the indirect measures, chapter 2 describes the development and content of the instruments used by teachers to rate children's academic and social skills as well as the instrument used by children to rate their own academic ability and interest and their behavior and relationships with peers. Chapter 3 contains an overview of item response theory (IRT) procedures used to scale the 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 third grade, and chapter 5 describes their role in longitudinal measurement. Chapter 6 describes the development and psychometric characteristics of the teacher indirect cognitive and social rating scale measures and the Self-Description Questionnaire administered to sampled children.

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 spring-kindergarten (round 2), fall- and spring-first grade (rounds 3 and 4, respectively), and spring-third grade (round 5). The third round (fall-first grade) was a subsample of about 30 percent of the kindergarten schools. The fifth round of data collection described in this report took place in spring 2002, when approximately 89 percent of the children were in third grade. The direct cognitive assessments were conducted in all five rounds of data collection, while the indirect cognitive and socioemotional measures were collected from teachers in rounds 1, 2, 4, and 5 (fall- and spring-kindergarten, spring-first grade, spring-third grade), and from parents in rounds 1, 2, and 4. In round 5, 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 *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), User's Manual for the ECLS-K Third Grade Public-Use Data File and Electronic Codebook* (NCES 2004–001).

Sample counts, completion rates, psychometric characteristics, and score statistics for the third 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 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 codebook and data file users' manuals. 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 Kindergarten Through First Grade* (NCES 2002–05).

This page is intentionally left blank.

2. DESIGN AND DEVELOPMENT OF THE ASSESSMENT INSTRUMENTS

The ECLS-K assessment instruments were designed to measure children's academic and social development during the kindergarten through fifth 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 and approaches to learning are reported in the social rating scales derived from teachers' observations in the school setting, as well as in children's self-reports. This chapter documents the design and development of the assessment measures used in the fifth round of data collection, when most of the ECLS-K children were in third grade.

The National Center for Education Statistics (NCES) and contractor staff assembled school curriculum specialists, teachers, and academicians 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 make efficient use of resources while minimizing burden on teachers and students.

The third grade direct cognitive assessments built on the structure established in the kindergarten and first grade (K-1) rounds of data collection. Individually administered assessments were conducted for the direct cognitive measures, while teachers provided indirect reports of children's academic skills, attitudes, and behaviors. A questionnaire eliciting children's academic and behavioral self-ratings was introduced in third grade. The third grade assessment battery differed from that of K-1 in several important respects:

- **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. Since no freshening of the sample occurred in spring 2002, the number of sampled children who might still lack English proficiency two years later, in third grade, was assumed to be so small that the language screening

assessment would be unnecessary. Therefore, an English language screener was not administered in the third grade data collection.

- **Changes in the content and format of the direct and indirect cognitive assessment instruments:** New reading and mathematics assessment forms were developed for the third grade. A science assessment replaced the direct cognitive assessment of general knowledge that had been used in kindergarten and first grade, while K-1 teacher ratings of general knowledge were separated into science and social studies sections. Assessment formats were similar to the earlier rounds, but some modifications were made to accommodate the content of the questions. A Spanish translation of the mathematics assessment, used in kindergarten and first grade, was assumed to be unnecessary for third grade.[1] Additional scores were defined that targeted third grade skills. Details of these changes are described in sections 2.1 and 2.2.

- **Self-Description Questionnaire:** In third grade, for the first time, children were asked to rate their own academic competence and interest and to report on their relationships with peers. See section 2.3 for more details.

- **No parent questionnaire items on children's social behaviors:** Parents' ratings of children's behavior and social skills had been collected during kindergarten and first grade rounds. These ratings were deleted from parent information collected in third grade for several reasons: age appropriateness of the instrument, 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 as a baseline measure and was not repeated in subsequent kindergarten, first grade, or third grade data collections.

Another change in the longitudinal design of ECLS-K was the elimination of the second grade round 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.

2.1 Direct Cognitive Assessment

The child development and primary education experts consulted by project staff during the design phase of the ECLS-K recommended that the knowledge and skills assessed by the ECLS-K tests should represent the typical and important cognitive goals of elementary schools' curricula. Therefore, the subject-matter domains of language and literacy skills (referred to hereafter simply as "reading" for the

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

direct cognitive assessment), mathematics, and science were selected for the third grade direct cognitive battery. Time constraints and concern about burden on children as well as differences in social studies curricula throughout the states led to a decision not to include a social studies assessment in the direct cognitive battery. 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 throughout the whole span of the study. The longitudinal design of the study required the development of vertical scales in reading and mathematics to support 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 students and teachers determined the kinds of test items that could be used, as well as the structure of the tests. The total time available to test each student in all three domains combined had to be less than 75 minutes, on average, in third grade. This 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 most difficult forms within the same sets. Because children's academic skills in third grade could be expected to have advanced beyond the levels covered by the K-1 assessments, a new set of assessment instruments was developed for the third grade. Some of the K-1 test items were retained in the third grade forms 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 direct cognitive science assessment in third grade. There is no longitudinal scale for measuring gains in science prior to third grade, because the third grade science assessment is not comparable to the K-1 general knowledge assessment. The substitution of a science assessment for the general knowledge assessment from third grade onward means that gains in

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

The format of the third grade assessment was similar to that of prior rounds, with some changes to accommodate the more advanced level of the questions. As in K-1, an assessor presented the questions to the child and entered responses into a computer for each individually administered assessment. A workbook of one to seven questions that required computations or written responses was added to the third grade mathematics assessment. The reading assessment in third grade was administered in booklet format (instead of the easel used in K-1) to accommodate the length of the reading passages used in the assessment.

Kindergarten and first grade children whose English language skills were not sufficiently advanced to be assessed in English and who were Spanish speakers were administered a Spanish translation of the ECLS-K mathematics assessment. No such translation was used for the reading and general knowledge assessments, which were too language- and culture-dependent to yield comparable measurement. More than two-thirds of the children who received the Spanish mathematics assessment in fall-kindergarten were able to take the English version by spring-first grade. The third grade battery was administered entirely in English.

The types of scores reported for the third grade direct cognitive assessments are similar to those for kindergarten and first grade, with some modifications for scores representing both broad-based measures and targeted skills. Assessment scores were rescaled for third grade, and several new scores were added. 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 third grade. As a result, scores in the public-use file for the kindergarten and first grade rounds should not be compared with recalibrated scores in the kindergarten through third grade public-use file. The kindergarten and first grade scores have been rescaled and appear in the kindergarten through third grade file so that comparisons are possible. New targeted scores based on clusters of third grade reading and science items are reported, and new proficiency levels are defined that correspond to grade-appropriate skills in reading and mathematics. Descriptions of scores appear in chapter 4.

2.1.1 Individually Administered Adaptive Tests

During the background review prior to the kindergarten year, the project staff, which included experts in child development, primary education, and testing methodology, made the recommendation that the direct cognitive measures be administered individually to each sampled child. 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 is most appropriate for his or her level of achievement.

The development of a vertical scale that must span kindergarten to fifth 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 in which 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 impact 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 discussion of IRT may be found in chapter 3, and notes on the ECLS-K longitudinal scales in chapter 5.

It is for these reasons that the ECLS-K uses individually administered adaptive tests. A review of commercially available tests indicated that there were no "off-the-shelf" tests that matched the domain requirements and were also both individually administered and adaptive. Individual administration of assessments was retained in third grade, even though children might have been able to cope with paper and pencil test forms at this time. The success of the adaptive approach in kindergarten and first grade in optimizing measurement characteristics for a diverse sample of children suggested its use in the later rounds as well. A change to group administration was considered for third grade, but rejected because it would have been difficult to administer given the two-stage adaptive structure of the assessments.

In the kindergarten and first grade rounds, a concern was expressed that the individual mode of administration may have contributed unwanted sources of variance to the children's performance in the direct cognitive measures. Unlike group administrations, which in theory are more easily standardized, variance attributable to individual administrators might affect children's scores. A multilevel analysis of fall-kindergarten and spring-first grade data found only a very small interviewer effect of about 1 to 3 percent of variance. A team leader effect could not be isolated, because it was almost completely confounded with primary sampling unit. Analysis of interviewer effect was not carried out for the third grade data for two reasons. First, the effect in K-1 was about twice as large for the general knowledge assessment (which was not used in third grade) than for reading or mathematics. Second, the effect found was so small that it was inconsequential. Refer to the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05) for more details on the analysis of interviewer effects.

2.1.2 The ECLS-K Frameworks

The ECLS-K is 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, two years after the design phase, and will continue until 2004.

The ECLS-K assessment frameworks were derived from multiple sources. A review of national and state performance standards, comparison with state and commercial assessments, and the judgments of curriculum experts and teachers all provided input to the ECLS-K test specifications. For the third through fifth grade assessments, national and state performance standards in each of the domains were examined. The scope and sequence of materials from state assessments, as well as from major publishers, were also considered.

Some of the ECLS-K panel consultants had been instrumental in developing the fourth grade National Assessment of Educational Progress (NAEP) content and process frameworks for reading, mathematics, science and social studies. 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, 1994, and 1996 frameworks were particularly useful as models for the third and fifth grade ECLS-K assessments since they define appropriate sets of skills and understandings at fourth grade. The resulting ECLS-K frameworks are similar to the NAEP fourth grade frameworks, with grade-appropriate modifications as well as 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 "best practices," as well as those that are widely taught. In contrast, the ECLS-K examines the full range of practices rather than concentrating on best practices. Nonetheless, 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, which 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 curricular content is not available in most subject areas so the study relied mainly on the advice of curriculum specialists and people with extensive teaching and administrative experience in elementary schools and on the standards published by states and national professional organizations. The overall testing time for each child was expected to consist of comparable time allotted for reading and mathematics, with a lesser amount of time allocated for the science assessment. It is important to keep in mind that some content strands can be assessed more quickly than other areas. For example, many single-word decoding 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 to fifth 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 and fifth grades, many items tap more than one area. For example, solving a mathematics problem may require understanding of number concepts as well as skill in interpreting data.

Table 2-1. Reading longitudinal test specifications for kindergarten through fifth grade: School years 1998–2004

Grade levels	Total	Basic skills	Vocabulary	Reading comprehension skills			
				Initial understanding	Developing interpretation	Personal reflection	Critical stance
			Percent of testing time				
Kindergarten	100	40	10	10	25	10	5
First grade	100	40	10	10	25	10	5
			Percent of test items				
Third grade	100	15	10	15	30	15	15
Fifth grade	100	10	10	15	30	15	20

NOTE: The column headings are identical to the NAEP 1994 Reading Framework categories, with the addition of Basic Skills and Vocabulary. Basic Skills include familiarity with print, recognition of letters and phonemes, and decoding. Initial understanding requires readers to provide an initial impression or global understanding of what they have read. Developing interpretation requires readers to extend their initial impressions to develop a more complete understanding of what was read. Personal reflection and response requires readers to connect knowledge from the text with their own personal background knowledge. The focus here is relating text to personal knowledge. Demonstrating a critical stance requires the reader to stand apart from the text and consider it objectively.

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, and spring 2002.

Table 2-2. Mathematics longitudinal test specifications for kindergarten through fifth grade: School years 1998–2004

Grade levels	Total	Content strands				
		Number sense, properties, and operations	Measurement	Geometry and spatial sense	Data analysis, statistics and probability	Patterns, algebra, and functions
		Percent of 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

NOTE: The content strands are identical to those used in the "Mathematics Framework for the 1996 National Assessment of Educational Progress (NAEP)," (National Assessment Governing Board, 1996a). The content strand item targets for the third and fifth grades match the NAEP fourth grade recommendations for the minimum number of "Number Sense" items and the maximum numbers for the other strands.
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, and spring 2002..

Table 2-3. Science longitudinal test specifications, in percent of test items, for third (spring 2002) and fifth grade (spring 2004)

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

NOTE: The ECLS-K science expert panel developed the column categories 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 total 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), spring2002, and spring 2004.

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, 1994a). The NAEP framework is defined in terms of four types of reading comprehension skills:

- **Initial understanding** requires readers to provide an initial impression or global understanding of what they have read. Identifying the main point of a passage and identifying the specific points that were drawn on by the reader to construct that main point would be included in this category.

- **Developing interpretation** requires readers to extend their initial impressions to develop a more complete understanding of what was read. It involves the linking of information across parts of the text, as well as focusing on specific information.

- **Personal reflection and response** requires readers to connect knowledge from the text with their own personal background knowledge. Personal background knowledge in this sense includes both reflective self-understanding, as well as the broad range of knowledge about people, events, and objects that children bring to the task of interpreting texts.

- **Demonstrating a critical stance** requires the reader to stand apart from the text and consider it objectively. This would include questions asking about the adequacy of evidence used to make a point or the consistency of someone's reasoning in taking a particular value stance. In kindergarten and first grade, some questions about unrealistic stories were asked to assess the child's notion of "real vs. imaginary." Such story types allow us to get information on critical skills as early as kindergarten. Third grade critical stance items might assess children's understanding of literary devices or the author's intention.

Because the NAEP framework begins with fourth grade, it had to be modified for ECLS-K to accommodate adequately the basic skills typically emphasized beginning in kindergarten. 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. It is also important to note that the ECLS-K asks teachers to provide information on each sampled child's writing abilities each year and on the kinds of activities they use in their classrooms to promote writing skills with the use of the Academic Rating Scale (see chapter 6 in this report).

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 National Assessment of Educational Progress (National Assessment Governing Board [NAGB], 1996a), which is in turn derived from the curriculum standards from the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics [NCTM] (1989). The content strands represented by the column categories in table 2-2 are defined as follows (these correspond closely to NAGB [1996a] definitions for most strands):

- **Number sense, properties, and operations.** This refers to children's understanding of numbers (whole numbers, fractions, decimals, and integers), operations, and estimation, and their application to real-world situations. Children are expected to demonstrate an understanding of numerical relationships as expressed in ratios, proportions, and percentages. This strand also includes understanding properties of numbers and operations, ability to generalize from numerical patterns, and verifying results.

- **Measurement.** Measurement skills include choosing a measurement unit, comparing the unit to the measurement object, and reporting the results of a measurement task. It includes items assessing children's understanding of concepts of time, money, temperature, length, perimeter, area, mass, and weight.

- **Geometry and spatial Sense.** Skills included in this content area extend from simple identification of geometric shapes to transformations and combinations of those shapes. The emphasis of the ECLS-K is on informal constructions rather than the traditional formal proofs that are usually taught in later grades.

- **Data analysis, statistics, and probability.** This includes the skills of collecting, organizing, reading, and representing data. Children are asked to describe patterns in the data or make inferences or draw conclusions based on the data. Probability refers to making judgments about the likelihood of something occurring based on information collected on past occurrences of the event in question. Students answer questions about chance situations, such as the likelihood of selecting a marble of a particular color in a blind draw when the numbers of marbles of different colors are known.

- **Patterns, algebra, and functions.** Consistent with the NCTM kindergarten to fourth grade curriculum standards, the ECLS-K framework groups pattern recognition together with algebra and functions. Patterns refers to the ability to recognize, create, explain, generalize, and extend patterns and sequences. In the kindergarten test, the items included in this category entirely consist of pattern recognition items. As one moves up to the subsequent grades, algebra and function items are added. Algebra refers to the techniques of identifying solutions to equations with one or more missing pieces or variables. This includes representing quantities and simple relationships among variables in graphical terms. It should be noted that while pattern recognition is heavily emphasized in kindergarten and even first grade classrooms, the proposed framework tends to de-emphasize the assessment allocation since it is not clear what to expect with reference to longitudinal trends in this skill area.

The number sense, properties, and operations content strand represents the dominant emphasis of elementary school mathematics. Additional discussion of the adaptation of the NAEP mathematics framework to ECLS-K, and an appendix listing the NCTM curriculum standards, may be found in the *ECLS-K Psychometric Report for Kindergarten Through the First Grade* (NCES 2002–05).

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 grade. No direct measurement of social studies knowledge was included in third grade, although teacher ratings of children's proficiency in social studies, along with the other subject areas, were collected and are described in chapter 6. For a discussion of the design and specifications of the K-1 general knowledge test, refer to the *ECLS-K Psychometric Report for Kindergarten Through the First Grade* (NCES 2002–05).

The test specifications for third and fifth grade science (table 2-3) were developed largely from recommendations of the ECLS-K advisory group. Similar to the 1996 NAEP Science Framework (NAGB, 1996b), 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 will be more on the adequacy of accounts than the grasp of discrete facts, particularly as the children move 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 the earth's composition, process, environments, and history, focusing on the solid earth and its interactions with air and water. The content to be assessed in earth science centers on objects (soil, minerals, rocks, fossils, rain, clouds, the sun and moon), as well as processes and events that are relatively accessible or visible. Examples of processes are erosion and deposition, and weather and climate; events include volcanic eruptions, earthquakes, and storms. Space science in the early elementary grades is usually concerned with the relationships between earth and other bodies in space (e.g., patterns of night and day and the seasons of the year, phases of the moon).

- **Physical science** includes matter and its transformations, energy and its transformations, and the motion of light, sound, and physical objects. Physical science concepts in the elementary grades include the physical and chemical transformations of matter such as liquids and solids, and the conduction of heat, sound, and electrical energy.

- **Life science** is devoted to understanding and explaining the nature and diversity of life and living things. The major concepts to be assessed relate to interdependence, adaptation, ecology, and health and the human body.

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.

2.1.3 Field Testing of Direct Cognitive Items

Preliminary pilot testing of assessment items was carried out for second through fifth grades in spring, 1999. Relatively small samples of children participated in the pilot tests, and relatively large numbers of test questions were tried out. Both multiple choice and open-ended items were used in each content domain. Items were revised on the basis of pilot test experience, and sets of questions were selected for a full-scale field test in spring 2000. At this point, the timing of the next round of ECLS-K data collection had not yet been determined, so second graders were included in the field test sample along with third graders, in the event that a second grade ECLS-K round might prove to be feasible. The field test results, in turn, were used to guide the revision and selection of items for the third grade assessments for the longitudinal sample.

2.1.3.1 Field Test Design

Preliminary pilot testing of items. Pools of test items in each of the content domains were developed for second through fifth grades. Items were chosen to extend the longitudinal scales initiated in kindergarten and first grade, with grade-appropriate changes in content and format. The majority of reading items for second through fifth grades tapped reading comprehension rather than basic skills. In mathematics, increased emphasis was placed on problem solving. Both of these areas made expanded use of open-ended items, and in both, children were asked to provide some of their answers on worksheets instead of orally. Some of the reading passages on which test questions were based were taken from published sources, while others were written for the ECLS-K. All of the mathematics and science questions were prepared by the ECLS-K item writers. Some utilized photographs or diagrams from published sources.

Test items were reviewed by elementary school curriculum specialists for difficulty, appropriateness of content, and relevance to the test framework. In addition, items were reviewed for sensitivity issues related to population subgroups. Items that passed these content, construct, and

sensitivity screenings were assembled into pairs of booklets for preliminary pilot testing in spring 1999. Approximately 120 to 150 items in each content area were distributed among two reading, two mathematics, and two science forms within each of the four grades. Each pilot test form in each grade, second through fifth, was administered to about 50 children. The results of the pilot testing were used to select and revise test questions for use in full-scale field tests of second and third graders in spring 2000, with field testing of fourth and fifth graders deferred until spring 2002.

Field test issues. The operational feasibility of the individualized two-stage assessment procedure with "on-time" scoring of the routing test had been established in the ECLS-K kindergarten and first grade rounds. These data collections had also satisfactorily demonstrated young children's ability to maintain the necessary attention span and to complete the assessments without signs of discomfort or distress. The field test for second and third grade 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 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, the Woodcock-McGrew-Werder Mini-Battery of Achievement (MBA; Woodcock, McGrew, and Werder, 1994). MBA subtests measuring letter and word identification, vocabulary, and comprehension were used for validation of the ECLS-K reading item pool, while validation of the mathematics pool was based on scores on the MBA Calculation and the Reasoning and Concepts subtests. The MBA subtests were administered according to standard procedures specified by the publisher.

Spring 2000 field test. About 120 to 130 questions in each of the content areas, i.e., reading, mathematics and science, were field tested. The items within each of the content areas were divided into two parallel sets of items, A and B. Six booklets, each containing three subtests, were created. Two of the three subtests in each booklet contained sets of test questions, in two different content areas. The third section of each booklet contained either the reading or mathematics MBA subtests that were used to validate the reading and mathematics item pools, or a set of academic and social skills questions being evaluated for use in a self-description questionnaire. Each A or B set of content area items appeared in one test booklet as the first cognitive set, and in another as the second, so that possible practice effects or fatigue effects would be balanced. For each of the six student test booklets, a corresponding examiner booklet contained the instructions for administering and scoring each test item. Booklet covers were color-coded for ease in matching the examiner to the student forms. Table 2-4 shows the subtests in each of the field test forms. The number of items (or separately scored item parts) is shown in parentheses after the name of the test section.

Table 2-4. Distribution of questions from the ECLS-K field test pool and the Mini-Battery of Achievement (MBA) mathematics and reading subtests in field test forms, by section: Spring 2000 field test

Field test form	Section 1	Section 2	Section 3
1: Red	Mathematics A (65)	Reading A (60)	Mathematics MBA (29,50)
2: Orange	Reading A (60)	Mathematics A (65)	Reading MBA (28,22,23)
3: Yellow	Socioemotional A (40)	Science A (60)	Mathematics B (63)
4: Green	Mathematics B (63)	Science B (64)	Mathematics MBA (29,50)
5: Blue	Reading B (60)	Science A (60)	Reading MBA (28,22,23)
6: Purple	Socioemotional B (41)	Science B (64)	Reading B (60)

NOTE: Number of items in each form shown in parentheses. The Mini-Battery of Achievement (MBA) Mathematics Part 3A. Calculation and Part 3B. Reasoning & Concepts and Reading Part A. Identification, Part B. Vocabulary, and Part C. Comprehension.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2000 field test.

About 1,800 children, evenly divided between second and third graders, participated in the field test of cognitive items in spring 2000. Each child was administered one of the six booklets. Spiralling the forms among test takers resulted in approximately 600 observations on each test question, about half of which came from second graders and half from third graders.

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 that could be revised, and detection of Differential Item Functioning (DIF) with respect to population subgroups. Validation of the ECLS-K reading and mathematics field test item pools was carried out by correlating field test ability estimates with MBA reading and mathematics test scores. Operational issues examined included timing, completion rates, and order effects. 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 third 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 second and third 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 right/wrong on each item. Distractor analysis consisted of

evaluating statistics on the percent 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 in need of revision, for example, questions that might have more than one potentially correct answer, or 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 other 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 test items. 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 trace plots were inspected for indications of lack of fit. The item trace plots identified the residuals separately for second and third graders, so suitability for each grade could be evaluated. A relatively small percentage of items exhibited overall lack of fit and were removed from consideration for the third grade battery. Examination of item plots for the poorer fitting items, along with the distracter analysis from the classical item statistics, can suggest possible revisions that might correct a flawed item. In some cases modifications to the response options could be made, and the item kept in the pool. Attempts to modify and retain flawed items were particularly important for items that represented one of the more difficult-to-fill cells in the framework classifications.

IRT-based estimates of ability distributions provided a basis for the selection of target difficulty ranges for the third 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 whose difficulty was matched to the ability levels that could be expected in the full-scale third 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 third grade forms.

Examination of the field test results confirmed that the absence of the second grade data collection from the longitudinal design might seriously impact the ECLS-K objectives. Section 2.1.5 discusses field test findings related to this issue, and the bridge study designed to support longitudinal measurement of gain.

Cognitive test items were checked for DIF for males compared with females and for Black students compared with White students. There were too few Hispanic and Asian children in the field test sample for DIF analyses to be carried out for these groups. It is not necessarily expected that different subgroups of students will have the same average performance on a set of items. But when students 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 student'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 students) with a matched reference group (e.g., males or White children). 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. 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. However, the 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. DIF analysis of field test items resulted in 5 mathematics and 5 science items being deleted from consideration for the third grade assessments; no reading items were affected.

Correlations between total scores on the MBA construct validation instrument and the IRT ability estimate (theta) from the field test items were computed for reading and mathematics. The high correlations (.83 for reading, .84 for mathematics) indicate strong similarity between the constructs being measured.

Factor analysis of field test sections was carried out for each of the subject areas. The ratio of first to second eigenvalues was high (more than 4:1 for reading and science and 6:1 for mathematics), suggesting a strong single factor underlying test performance. This finding supports the use of IRT scaling. Attempts to identify additional distinct factors within each subject test resulted in factors related to item difficulty, but not to content strands. The implications of this are relevant to the discussion of item selection in section 2.1.4.

Operational issues. Findings from both quantitative and qualitative analysis of field test data answered questions related to practical and administrative issues, such as timing, fatigue, and cooperation. Analysis of item statistics showed that the time allotted for each test section was sufficient for most children to answer all or nearly all of the field test items. Item nonresponse rates were low: relatively few children either omitted items while answering subsequent questions or failed to reach the end of the test sections.

The booklet design described earlier, with each test form appearing both early and late in a testing session, permitted analysis of order effects. Comparison of statistics for sets of items given early in the testing session with the same items given in a later position showed only a small number of significant differences, about what would be expected by chance alone. This suggests that neither a practice effect (better performance toward the end of the test) nor a fatigue effect (a drop in performance) was operating, and that the one hour of testing time was not burdensome for most children.

For the reading sections, field test assessors recorded whether each child read passages aloud, silently, or in a mixture of the two modes. The majority of children, more than two-thirds over all, consistently read silently, while very few combined the two modes. Item score comparisons and timing comparisons were carried out for children who read the reading passages aloud compared with those who read silently. Performance differences were substantial for the two modes: the percent of correct answers for most test questions was higher for silent readers than for children who read aloud. Children who read aloud also took somewhat longer to finish the tasks. Since the ECLS-K tests were designed to be unspeeded power tests, this finding suggested that children should not be directed to read passages silently, but might maximize their comprehension if allowed to read in whatever mode they chose.

Field test assessors participated in debriefing sessions following the spring 2000 field test administration. They provided information on the children's reactions to test questions as well as suggestions on revisions of items that might improve item performance. They reported that most of the children were interested and cooperative. The assessors made numerous suggestions about item content, presentation, and scoring. Comments related to performance (such as reports that children found an item too difficult, or confusing or ambiguous) were, in general, corroborated by analysis of the field test data. The assessors' suggestions were taken into consideration in the selection and revision of items for the third grade assessments.

2.1.4 Third Grade Test Forms

The third grade assessments were designed to support measurement of the reading, mathematics, and science domains as accurately as possible, both at all levels of ability found within the ECLS-K third 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, had no DIF problems related to subgroup membership, and showed some increase in percent correct between second and third graders. Framework specifications, and practical issues such as timing and scoreability of items, placed additional constraints on assessment design. Design of the test forms required some compromises due to competing objectives.

2.1.4.1 Item Quality and Reliability

In order to contribute useful information about children's skill levels, test items selected for the final forms should ideally have high *r*-biserials (.40 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 K-1 assessments, or links to a selected reading passage. Reliability of the measure is improved by administering the maximum number of items possible in the time available, while maintaining high internal consistency. Items found to have DIF for population subgroups were deleted from the item pool.

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 third graders was used to determine item difficulty objectives such that the middle-difficulty form would be suitable for approximately the middle half of third grade test takers, while the low and high second-stage forms would each be taken by about a quarter of the children. Thus, the target difficulties for the majority of the second-stage middle form items were selected to fall within two-thirds of a standard deviation above and below the mean third grade ability estimate, corresponding to 50 percent of the distribution. The low and high second-stage forms consisted primarily of easier and harder items, respectively. The low form items ranged from about two standard deviations to about two-thirds of a standard deviation below the third grade mean, overlapping with some of the easier items in the middle form. Each high second-stage form began with items overlapping the hardest middle form items, at about two-thirds of a standard deviation above the mean, and ranged up to two standard deviations above the mean. 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 third grade assessment forms were designed to have enough easy items that distinctions can be made at the low end of the ability range, and enough hard items to accurately measure the most skilled students. In order 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 contains some items with difficulty levels that extend beyond the target ability range, at both the high and low end. This design feature serves two purposes. First, it provides 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 improves measurement properties for children whose achievement level is very near a routing cut point. There is 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 is too easy or too hard. For example, a child whose ability level is half a standard deviation below the mean (i.e., near the low end of the middle ability range) might miss a few routing test items and be assigned to the low second-stage form. Accuracy of measurement in this situation is supported by the overlap of some of the hardest low form items with the easiest middle form items.

2.1.4.4 Longitudinal Score Scale

Measurement of gain over time requires a longitudinal score scale. The challenge for 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 second-stage forms. The third grade assessments used the same overlapping two-stage design, but with more advanced sets of items. Putting K-1 and third grade scores on a common scale required common items

shared between the K-1 and third grade assessments. Items were selected from the K-1 assessments (22 in reading, and 14 in mathematics) to provide the necessary link to third grade. Most of these common items were too difficult for the vast majority of first graders, and too easy for almost all of the third graders. The link between first and third grade relied heavily on the collection of test data from a sample of second graders to bridge the gap in ability distributions and stabilize item parameters. See sections 2.1.5 and 5.1 for details on the bridge study design and results.

2.1.4.5 Curriculum Relevance

Both second and third graders participated in the 2000 field test of cognitive items. Although there was no second grade round of data collection for the longitudinal sample, the second grade field test data did play a role in the design of the test forms for the third grade longitudinal sample. Analysis of field test data was carried out for both grades combined, as well as separately for grades 2 and 3. In selecting items for the third grade test forms, preference was given to items that showed the largest differences in percent correct between the second graders and third graders in the field test sample. Although the second and third graders in the field test were different children, not longitudinal measurements of the same children, items with the largest second to third grade differences in percent correct could be assumed to be strongly related to third grade curriculum. This inference was supported by the finding that not all items showed large differences. Many had close to the same percent correct for second grade and third grade field test participants, suggesting that their content was not emphasized in third grade curriculum materials.

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 K-1 scales, 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 four to six items or more 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

Table 2-5. Reading third grade framework targets and percent of assessment items: School year 2001–02

Percent of assessment items	Total	Basic skills	Vocabulary	Initial understanding	Developing interpretation	Personal reflection	Critical stance
Target	100	15	10	15	30	15	15
Actual	100	27	12	21	21	4	14

NOTE: Detail may not sum to total due to rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

Table 2-6. Mathematics third grade framework targets and percent of assessment items: School year 2001–02

Percent of assessment items	Total	Number sense, properties, and operations	Measurement	Geometry and spatial sense	Data analysis, statistics and probability	Patterns, algebra, and functions
Target	100	40	20	15	10	15
Actual	100	44	18	13	12	13

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

Table 2-7. Science third grade framework targets and percent of assessment items: School year 2001–02

Percent of assessment items	Total	Earth and space science	Physical science	Life science
Target	100	33	33	33
Actual	100	34	34	32

NOTE: Detail may not sum to total due to rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

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.

Items in the Basic Skills strand in reading were overrepresented in the third grade reading assessment primarily because of the objectives of linking the scale to K-1 and avoiding floor effects. The reading framework called for 40 percent of the assessment time in Basic Skills items in kindergarten and first grade, but only 15 percent of the items in third grade. A majority of the items needed for the first to third grade link were decoding items classified as Basic Skills, and these same items served to avoid floor effects for the lowest achieving third graders. Very few of these easy items would have been needed for the third grade forms if not for the need to establish a longitudinal scale. Although the strand was already overrepresented, additional difficult decoding items (also classified as Basic Skills) were selected because they filled gaps in the distribution of item difficulties. Similarly, Initial Understanding items were overrepresented in comparison to framework targets. Eight items in this category were retained for the K-1 link, while others were selected because they accompanied a selected reading passage.

Underrepresentation occurred for two of the reading strands, Personal Reflection and Developing Interpretation. There would have been barely enough field-tested Personal Reflection items available to meet the framework specifications, even if all of the field tested items in this strand were usable. However, three of the Personal Reflection items were rejected due to poor psychometric characteristics, and three others in the same category were deleted from the pool because of negative feedback from assessors combined with poor item statistics. The situation was similar for Developing Interpretation items.

Comparable reasons for deviations from framework targets were encountered for the mathematics assessment. The need to link K-1 to third grade, including retaining proficiency cluster items (see section 4.1), resulted in an overrepresentation of Number Sense/Properties/Operations items. Meanwhile geometry items were underrepresented, even after some exceptions to quality standards were made, primarily because of weak psychometric characteristics. No problems were encountered in selecting science items to match framework percentages, in large part because the constraint of linking to K-1 was not present for the new science assessment. Enough high quality science items were available for selection in each of the content strands.

These deviations from framework targets probably have relatively little impact on the measurement of the domain of interest, for two reasons. First, there is some ambiguity in the classification of items. Many if not most of the third grade reading and mathematics items had aspects of more than one content strand. 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. Even the Basic Skills decoding items were probably affected by children's mastery of vocabulary. 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. The second reason justifying some flexibility in matching target framework percentages was that the factor analysis mentioned in section 2.1.3.2 found each of the sets of field test items to be strongly single factor. Underlying factors related to the strand categories were not found for any of the subject areas. Therefore, it is likely that compromises in selecting items would not have a serious negative impact on measurement of the intended construct.

2.1.4.7 Practical Issues

The 75-minute time allocation for the third grade direct cognitive assessments was divided into 30 minutes each for reading and mathematics and 15 minutes for science. Analysis of field test timings showed that more time per item was needed for reading, with the extra time required for the reading passages, and for mathematics, which required problem solving, than for science questions. The sets of science items, consisting of short-answer questions, tended to go much more quickly. The number of items in each of the third 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 third graders, 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 three proposed second-stage forms. Crosstabulations 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. For example, if many of the children with simulated routing scores below 7 would be expected to receive below-chance scores on the middle difficulty item set, but few if any perfect scores on the low second-stage items, children in this routing score range would be assigned to the low form. This procedure was carried out rather than relying on cut points that approximated the planned 25-

50-25 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 cutting 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 third grade longitudinal sample. The success of the two-stage test design in achieving its goals is discussed there as well.

Table 2-8. Number of items in third grade test forms and routing test cut scores, by domain: School year 2001–02

Description	Reading	Mathematics	Science
Number of items per form			
Routing test	15	17	15
Low second-stage form	24	25 (23 scored)	20
Middle second-stage form	39 (38 scored)	24 (23 scored)	20
High second-stage form	42 (36 scored)	24 (23 scored)	20
Total number of items			
Third grade pool	88	77	62
K-1 Pool	92	64	†
Overlap between K-1 and third grade	22	13	†
Items in longitudinal scale (K-1 + third grade)	154	123	62
Routing test cut scores			
Route to low second-stage form	0–8		0–6
Route to middle second-stage form	9–12	7–11	7–11
Route to high second-stage form	13–15	12–17	12–

† Not applicable.
NOTE: The number of items in each third grade pool is less than the sum of the items in the test forms because there is some overlap of items across forms. Four third grade reading items were not scored because they proved to be too difficult to provide useful information, and two others because statistics were unsatisfactory. Two mathematics items were deleted from scoring because of differential item functioning (DIF) in the third grade sample, and two others because of poor fit to the item response theory (IRT) model. 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 2002.

Test administration procedures called for assessors to record children's selected response options for multiple choice questions, or a "1" (correct) or "2" (incorrect) for open ended items. Scoring protocols for the open ended items were provided to the assessors to ensure that assessors scored each response accurately and as objectively as possible. During debriefing sessions following the field test, assessors provided feedback on the adequacy of the scoring protocols. Their input contributed to revisions of scoring protocols, including clarifying ambiguous material and adding unanticipated responses received from field test children to the lists of correct or incorrect responses. A few items that assessors felt could not be scored objectively were deleted from item pools, if field test statistics (such as low *r-biserial*s) corroborated their reports.

Experts in each of the subject areas reviewed the proposed third grade forms for appropriateness of content and relevance to the assessment framework.

2.1.5 Bridge Sample

One of the critical goals of the ECLS-K is to measure children's growth in cognitive achievement across the early elementary school years. Due to budgetary constraints, data were not collected in 2000–01, when most of the sampled children were in second grade. The absence of second grade data presented a challenge for establishing longitudinal scales to link the first grade to third grade scores. Very few children answered the most difficult items in the spring-first grade data collection correctly. Third grade field test results indicated that these same items would be too easy for the vast majority of third graders. The ability levels of the highest scoring first graders overlapped with those of the lowest scoring third graders only in the tails of the distributions. This was particularly true for the reading assessment: by the end of first grade, most children had just begun to develop reading comprehension skills, while field test results indicated that by spring of third grade, most children were reading fluently. A similar pattern, with only slightly more overlap in first to third grade ability distributions, was found for mathematics. Without any second grade data, it would be difficult to place items reliably along the difficulty scale, making it impossible to accurately estimate cognitive gains from first to third grade. Based on the field test results, a bridge study was proposed that would fill in data points that lie between the preponderance of the first and third grade ability levels so that stable item parameter estimates could be calculated that would support the measurement of gain. Reading and mathematics assessment data collected from a relatively small sample of second graders would serve as a bridge to link the first and third grade rounds. Such a bridge sample would not need to consist of ECLS-K longitudinal sample members, nor would it need to be nationally representative. It would merely need to supply information on the performance of the ECLS-K assessment instruments at achievement levels typical of second graders.

The initial plans for the second grade bridge sample called for the design of a set of second grade assessment instruments, with items overlapping the K-1 and third grade assessments. Statistics obtained for the field test second graders showed a substantial amount of overlap between the second and third grade ability distributions for reading and mathematics. Most of the test items suitable for second graders had already been selected for the third grade assessments, primarily for the reading and

mathematics routing tests and low second-stage forms. They contained numerous items drawn from the K-1 forms. As a result, no new instruments were needed for the second grade bridge sample. The third grade forms and procedures were used, with the expectation that a majority of the second graders would route to the low second-stage reading and mathematics forms. Results of the bridge study are presented in section 5.1.

2.2 Indirect Measures: Teacher Ratings

Teachers of ECLS-K children received two questionnaires (A and B) asking about their background, training, and classroom practices. They also received a third questionnaire (C) which asked the teacher to rate each ECLS-K child on sets of academic and behavioral measures. The following two sections describe the indirect assessments that the teachers were to complete in third questionnaire.

2.2.1 Academic Rating Scale

The Academic Rating Scale (ARS) indirect cognitive measures were developed for the ECLS-K to measure teachers' evaluations of students' academic achievement in four domains: language and literacy (reading and writing), mathematical thinking, science, and social studies. The ARS was designed both to overlap and to augment the information gathered through the direct cognitive assessment battery. Although three of the four rating scales measure children's skills and behaviors within the same broad curricular domains as the direct measures, some of the constructs they were designed to measure differ in significant ways. Most importantly, the ARS includes items designed to measure both the process and products of children's learning in school, whereas the direct cognitive battery assesses only the products of children's achievement. The scope of curricular content represented in the indirect measures was designed to be broader than the content represented on the direct cognitive measures. The direct cognitive battery was less able to measure the process of children's thinking, including the strategies they use to read, solve math problems, or investigate a scientific phenomenon. Due to format limitations, the direct cognitive battery was not able to assess writing skills, while time constraints and variations in curriculum standards precluded direct assessment of children's knowledge in social studies. On the ARS, teachers reported the ECLS-K children's knowledge and understanding in the following social studies content areas: civics, geography, history, culture, and economics.

Unlike the direct cognitive measures, which were designed to measure gain on a longitudinal vertical scale from kindergarten entry through the end of fifth grade, the ARS was targeted to a specific grade level. The questions ranged from criterion-referenced items (e.g., "Divides a 3-digit number by a 1-digit number") to others with a more norm-referenced point of view (e.g., "Uses various strategies to gain information" or "Communicates scientific information"). Each question includes examples that were meant to help teachers think of the range of situations in which the child might demonstrate similar skills and behaviors and to illustrate the level of proficiency a child should have reached in order to receive the highest rating. Teachers evaluating the children's skills were instructed to rate each child compared with the skills of other children of the same age or grade level.

The development of the indirect measures paralleled the development of the direct measures. A background review of the literature on the reliability and validity of teacher judgments of academic performance was conducted (see Meisels and Perry, 1996). National and state standards as well as the literature on the predictive validity of early skills were examined to develop the item pool. The following criteria were used in creating and selecting items for the ARS:

- Skills, knowledge, and behaviors that reflect the most recent state and national curriculum standards and guidelines;
- Variables identified in the literature as predictive of later achievement;
- Direct criterion-referenced items with high level of specificity that called for lower levels of teacher inference;
- Skills, knowledge, and behaviors that were easily observable by teachers;
- Items broad enough to allow for diverse populations of students to be evaluated fairly;
- Some items that overlapped with the content assessed through the direct cognitive battery;
- Some items that expanded the skills tested by the direct cognitive battery—particularly those that assess process skills that would be difficult to assess directly given the time constraints;
- Literacy items that targeted speaking, reading, and writing skills; and
- Items that reflected developmental change across time.

Teachers were to rate each child's skills, knowledge, and behaviors on a scale from "Not Yet" to "Proficient" (see exhibit 2-1). If a skill, knowledge, or behavior had not been introduced into the

classroom yet, the teacher coded that item as N/A (not applicable). In third grade, the classroom teacher most knowledgeable of the child's academic achievement in the four domains might not be the primary or homeroom teacher. The primary teacher was asked to forward the rating form to the teacher most knowledgeable of the particular domain to complete the ratings. The differences between the direct and indirect cognitive assessments and the scores available are described here. For a discussion of the content areas of the ARS, see the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), User's Manual for the ECLS-K Third-Grade Public-Use Data File and Electronic Code Book* (NCES 2004–001).

Exhibit 2-1. Academic Rating Scale response scale, third grade: School year 2001–02

1	Not yet:	Child *has not yet* demonstrated skill, knowledge, or behavior.
2	Beginning:	Child is *just beginning* to demonstrate skill, knowledge, or behavior but does so very inconsistently.
3	In progress:	Child demonstrates skill, knowledge, or behavior *with some regularity* but varies in level of competence.
4	Intermediate:	Child demonstrates skill, knowledge, or behavior *with increasing regularity and average competence* but is not completely proficient.
5	Proficient:	Child demonstrates skill, knowledge, or behavior *competently and consistently*.
	N/A:	Not applicable: Skill, knowledge, or behavior has *not been introduced* in classroom setting.

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

Teachers from both public and private schools and from different regions of the country and content experts familiar with the early grades reviewed the items and made recommendations. Items were then piloted and later field tested in order to gather statistical evidence of the appropriateness of the items for carrying out the overall assessment goals. The pilot testing indicated that the difficulty of the items needed to be increased in order to capture the range of abilities represented in third grade and to avoid a serious ceiling problem. The items were revised and the difficulty of the criteria in the exemplars increased before field testing. The items were field tested in the spring of 2000, at the same time as the field test of the direct cognitive assessments. Final items were chosen consistent with the item statistics and representativeness of the content.

2.2.2 Social Rating Scale

The Social Rating Scale (SRS) is an adaptation of the Social Skills Rating System (Gresham and Elliott, 1990). Teachers use a frequency scale (see exhibit 2-2) to report on how often the student demonstrates the social skill or behavior described. Factor analyses (both exploratory analyses and confirmatory factor analyses using LISREL) were used to confirm the scales. The 24 SRS items used in kindergarten and first grade were included in the third grade SRS, and two new items were added. For additional information on the SRS instrument, see section 6.1.2 of this report, sections 2.3.2 and 3.3 of the *ECLS-K User's Manual for the Third-Grade Restricted-Use Data File and Electronic Code Book* (NCES 2003–003) and the *ECLS-K Psychometric Report for the Kindergarten Through First Grade* (NCES 2002–05).

Exhibit 2-2. Social Rating Scale response scale, third grade: School year 2001–02

	Answer	Description
1.	Never	Student never exhibits this behavior.
2.	Sometimes	Student exhibits this behavior occasionally or sometimes.
3.	Often	Student exhibits this behavior regularly but not all the time.
4.	Very often	Student exhibits this behavior most of the time.
N/O.	No opportunity	No opportunity to observe this behavior.

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

A parent version of the SRS had been administered in the kindergarten and first grade years as part of a telephone or in-person survey. (See chapter 2 in the ECLS-K kindergarten and first grade user manuals for a more detailed description of the parent scales.) The factors on the parent SRS were similar to the teacher SRS; however, the items in the parent SRS were designed for the home environment and, thus, were not the same as the teacher items. It is also important to keep in mind that parents and teachers observe the children in very different environments. Results of the K-1 parent SRS are presented in the *ECLS-K Psychometric Report for the Kindergarten Through First Grade* (NCES 2002–05). A parent version of the SRS was not administered during the third grade parent interview.

2.3 Self-Description Questionnaire

For the first time in the ECLS-K, third grade students rated their perceived academic competence and social skills. The Self-Description Questionnaire (SDQ) 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, 1990) 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 field test: Reading, Mathematics, All School Subjects, and Peer. The response scale as well as several of the items were adapted for use, with permission, in the main study.

The original SDQ-I has some negatively worded items that were not scored, but were included in the instrument in order to break any response sets that might occur. 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 may interfere with learning. Items measuring both internalizing and externalizing problem behaviors were included. The internalizing problem behavior items included items tapping anxiety about school, sadness, and loneliness. The externalizing problem behavior items assessed acting out behaviors and attention problems. These scales also were tested in the spring 2000 field test.

After analyzing different combinations of responses, it was found that a 3- to 4-point response scale worked best. A 4-point scale offered the opportunity to get as much variance as possible within the ability of third graders to interpret the response choices. Children appeared hesitant to use the extreme negatively-laden ends of the response scale; thus the response choices used assessed degrees of truth rather than the degrees of truth and untruth used in the original SDQ-I: "not at all true," "a little bit true," "mostly true," or "very true." This also reduced the cognitive demand for the students.

The SDQ consisted of 42 statements, including self-ratings of children's competence and interest in reading, mathematics, and "all school subjects." The statements also included self-ratings of children's competence and popularity with peers and problem behaviors with which they might struggle. The following scales were used with ECLS-K students in the fifth round of data collection:

- **SDQ Reading** scale includes items about reading grades, the difficulty of reading work, and their interest in and enjoyment of reading. (8 items)

- **SDQ Mathematics** scale includes items about mathematics grades, the difficulty of mathematics work, and their interest in and enjoyment of mathematics. (8 items)

- **SDQ School** scale includes items about how well they do in "all school subjects" and their enjoyment of "all school subjects." (6 items)

- **SDQ Peer** scale includes items about how easily they make friends and get along with children as well as their perception of their popularity. (6 items)

- **SDQ Anger/Distractability** scale includes items about externalizing problem behaviors such as fighting and arguing "with other kids," talking and disturbing others, and problems with distractability. (7 items)

- **SDQ Sad/Lonely/Anxious** scale includes items about internalizing problem behaviors such as feeling "sad a lot of the time," feeling lonely, feeling ashamed of mistakes, and worrying about school and friendships. (7 items)

In addition to the change in response scale and the addition of problem behavior items, the following adaptations were made to the original SDQ-I.

- The word "marks" was changed to "grades" in items asking about their performance in reading, mathematics, and all school subjects.

- Items that were at similar difficulty levels were eliminated, when it did not affect the reliability, to decrease the number of items in the scale.

- Students had some difficulty understanding "look forward to…", so the wording was changed to "cannot wait to…."

- Items were added to decrease the number of children who rated themselves as very competent ("very true") on all items: "I can do very difficult math problems."; "I like reading chapter books."

For additional information about the changes made to the SDQ-I, see the field test report (Atkins-Burnett, Meisels, and Correnti, 2000) of the *Self-Description Questionnaire-I* for the second and third grades.

This page is intentionally left blank.

3. ANALYSIS METHODOLOGY

This chapter describes in detail the methodology used to carry out specialized procedures for psychometric analysis of ECLS-K third grade data. 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. A one-parameter (Rasch) model was employed for scoring teacher ratings with multiple categories. Differential item functioning (DIF) procedures identified test items that performed differently for subgroups of the population.

3.1 Overview: The Three-Parameter Model

Measuring the extent of cognitive gains at both the group and individual level requires that the various kindergarten through third 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 anchor 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 fifth 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 third and fifth grade) that taps materials normally found in the curriculum for higher grades, and that build on skills learned in earlier grades.

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

3.1.1 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, discriminating ability, and "guess-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. The point on the vertical probability axis corresponding to the height of the curve at a given value of theta is the estimated probability that a person of that ability level will answer the test item correctly. The shape of the curve is given by the following equation describing the probability of a correct answer on item *i* as

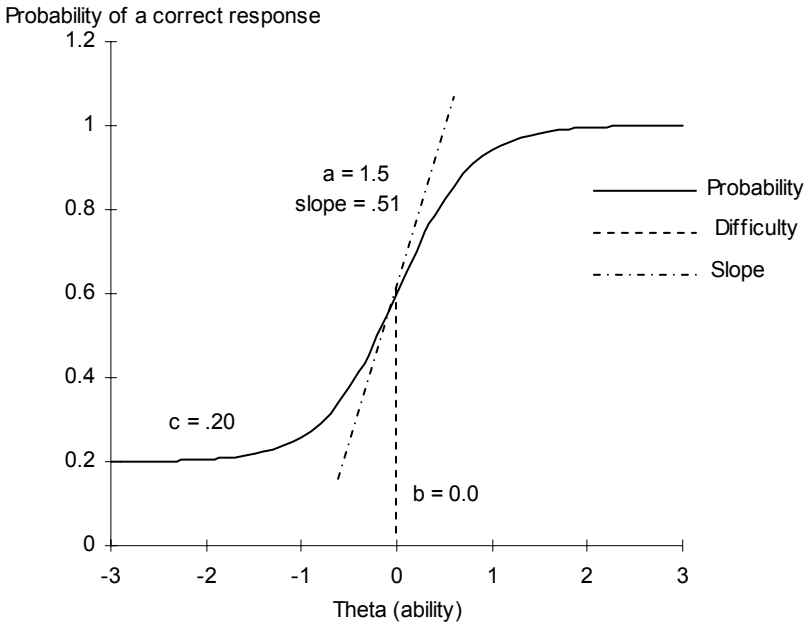
$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-1.702 * a_i(\theta - b_i)}}, \tag{3.1}$$

where θ = ability of the test taker;
a_i = discrimination of item i, or how well the item distinguishes between ability levels at a particular point;
b_i = difficulty of item i; and
c_i = "guessability" of item i.

The "c" parameter represents the probability that a test taker with very low ability will answer the item correctly. 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.

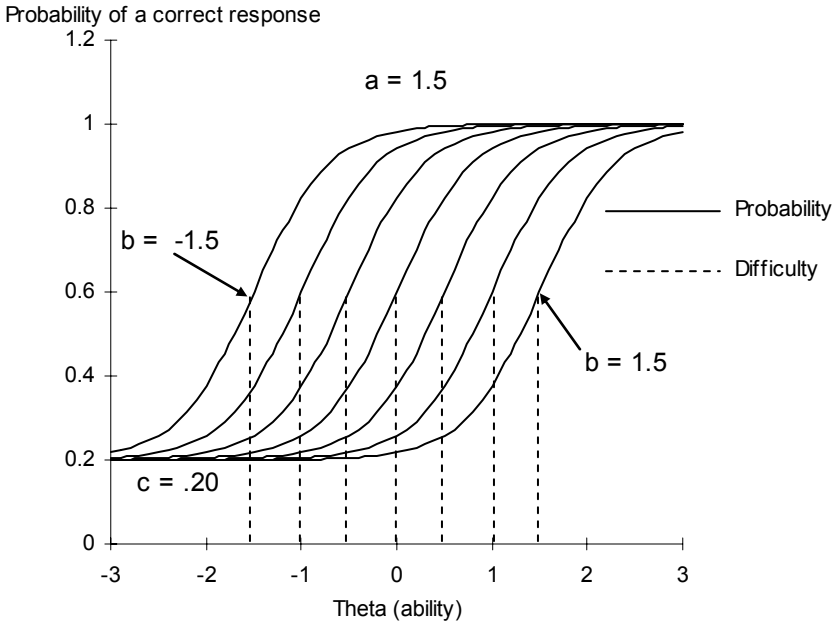
The IRT "b" parameters correspond to the difficulty of the items, represented by the horizontal axis in the ability metric. In figure 3-1, b = 0.0 means that test takers with θ = 0.0 have a probability of getting the answer correct that is equal to halfway between the guessing parameter and 1. In this 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 θ > b.

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



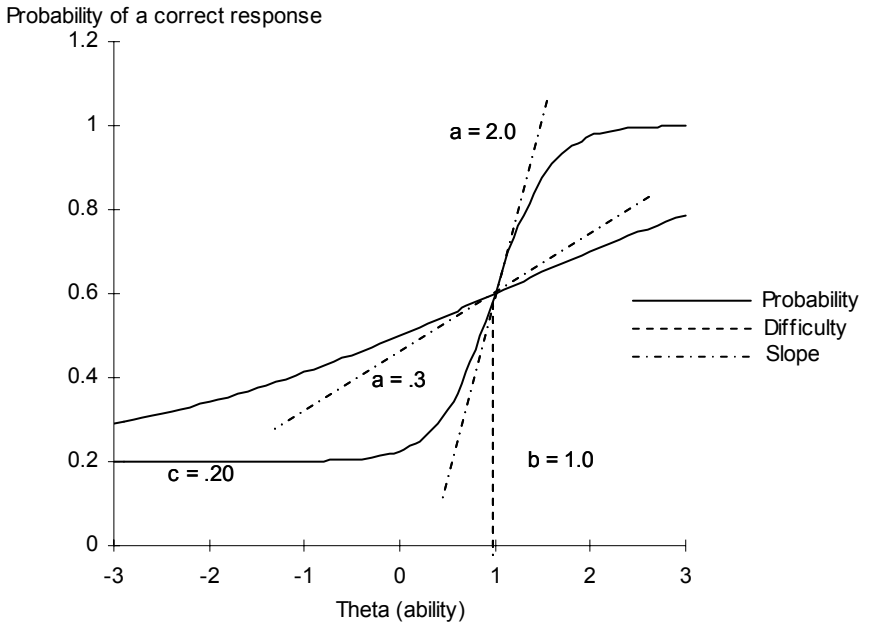
NOTE: a = parameter for discrimination; b = parameter for difficulty; and c = parameter for guessing. The discrimination parameter is proportional to the slope (tangent) of the function at the point of inflection.

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.

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.

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 likely to get it right at only about the guessing rate) from those of ability higher than the item "b," who are nearly certain 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 the 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 is 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" parameter, item difficulty, 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.

Once a pool of test items exists 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, 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 does not give credit for correct answers to hard items by low ability students. Omitted items are treated as if the examinee had guessed at random. Raw number-right scoring, in effect, treats omitted items as if they had been answered incorrectly. While this may be a reasonable assumption in a motivated test for older students, this may not always be the case in the ECLS-K, where behavioral or other factors may contribute to a child's inability to complete all items.

3.1.2 Item Response Theory Estimation Using PARSCALE

The PARSCALE (Muraki and Bock, 1991) computer program computes marginal maximum-likelihood estimates of IRT parameters that best fit the responses given by the test takers. The procedure calculates "a," "b," and "c" parameters for each test item, iterating until convergence within a specified level of accuracy is reached. Comparison of the IRT-estimated probability 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 curves and the actual data points 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, and third grade assessments had been completed (five rounds, counting fall and spring administrations in K-1) there are five 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 a sixth 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 was not enough data 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 were collected that could be used to stabilize the estimates for the middle- and then the high-level items. As each round of data became available, item responses were pooled and parameters re-estimated. The pooling of all time points and re-estimating 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, fall- and 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. The use of all data points over time is desirable because it can provide stable estimates of both the item traces and latent trait scores throughout the entire ability distribution. This procedure was used in the vertical equating that was carried out for 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 information about the ability distribution (i.e., 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 scores (i.e., correct answers to all items administered). 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 third grade than in first grade? Probably not. Pooling all available information, that is, pooling all item responses for all people at all time points, and recomputing all of the item parameters using Bayesian priors reflecting the ability distributions associated with each particular round, provides for an empirically based shrinkage 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 scores, which in turn allows for the potential of some gains even in the uppermost tail of the distribution. Each of the rounds of data collection in kindergarten through third grade is treated as a separate subpopulation with its own ability distribution. The amount of shrinkage is a function of the distance from the subgroup means and the relative reliability of the score being estimated. 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, 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 ceiling and floor 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 estimation is the same as that used in NAEP and described in some detail by Yamamoto and Mazzeo (1992, p. 158) as follows:

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

In equation (3.2), $P(x_{j:g} \mid \theta, \beta)$ is the conditional probability of observing a response vector $\mathrm{x}_{j:g}$ of person j from group g, given proficiency θ and vector of item parameters $\beta = (a_1, b_1, c_1, ..., a_k, b_k, c_k)$, 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 can be assumed known and held fixed during item parameter estimation or can be estimated concurrently with item parameters.

The $f_g(\theta)$ 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 $A_g(X_k)$ may be chosen to be the normalized density at point X_k (i.e., $A_g(X_k) = f_g(X_k) / \Sigma_k f_g(X_k)$).

Maximization of $L(\beta)$ is carried out by an application of an 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 $\hat{N}_k = \Sigma_g \hat{N}_{gk}$). These same

provisional estimates are also used to estimate an expected frequency of correct responses at each quadrature point for each group (denoted $\hat{r}_{gik}$), and over all groups (denoted $\hat{r}_{ik} = \sum_g \hat{r}_{gik}$). 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 six time points (the five longitudinal ECLS-K rounds plus the second grade bridge sample). The present version of the multiple group PARSCALE only saves 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 the latent variate, were obtained from the C-Group conditioning program, which uses the gaussian quadrature procedure. This variation is virtually equivalent to conditioning (e.g., see Mislevy et al., 1992) on a set of "dummy" variables defining which ability subpopulation an observation comes from. The one difference is that the group variances are not restricted to be equal as in the standard conditioning procedure.

Conditional independence is an assumption of all IRT models, but as Mislevy et al. point out, not likely to be generally true. 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, factor analyses were carried out on the field test forms to confirm that there was a large dominant factor underlying each content area. In addition, all item traces were inspected to ensure a good fit throughout the ability range. More importantly, estimated proportions correct by item by grade were also estimated in order to ensure that the IRT model was both reproducing the actual percent correct (P+) for each item and there was no

systematic bias in favor of any particular grade. Since the item parameters were estimated using a model that maximizes the goodness of fit across the rounds, one would not expect much difference. No systematic bias was found for any grade.

Appendices B-1 to B-3 list the IRT item parameters for the three subject areas. They also show the actual proportion correct for test takers who answered each item, the proportion correct predicted from the IRT model, and the difference. Note that the IRT difficulty ("b") parameters are not directly related to the proportion of correct answers, because different groups of children took different items. Actual and estimated percent correct are reported for each item based on only the cases with response data, while the IRT parameters place items along the continuum of the whole scale, regardless of which test form(s) contained each item. For example, an item that is very easy relative to the whole item pool (low b parameter) may not have a very high percent correct if it appeared only in the kindergarten through first grade (K-1) low second-stage form and was taken only by the lowest ability children. Conversely, a relatively hard item, with a high "b" parameter, might have a low percent correct if it appeared in a routing section taken by all children, but a high proportion of correct answers if it was taken only by children routed to the high second-stage form.

3.2 One-Parameter Item Response Theory: The Rasch Model

A Rasch model (Rasch, 1960) was used to estimate the scores on the Academic Rating Scale (ARS) described in chapter 6. In Rasch models (also called one-parameter logistic models), the log odds of the probability of a correct response are a function of the difference between the person's ability and the difficulty of the item. The item discrimination is held constant across the items, and there is no guessing parameter. Applying the Rasch model to the data allows one to construct invariant linear measures, estimate the accuracy of the measures (standard errors), and determine the degree to which these measures and their errors are confirmed in the data using the fit statistics (Wright, 1999). Like the three-parameter IRT models, the Rasch model assumes unidimensionality, that is, a single dimension is being measured.

The Rasch Rating Scale model (Wright and Masters, 1982) was used with the ARS data:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_k)]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} [\beta_n - (\delta_i + \tau_k)]}, \; x = 0.1. \ldots, m \qquad (3.3)$$

where

$\tau_0 = 0$ so that $\exp \sum_{j=0}^{0} \left[\beta_n - \left(\delta_i + \tau_j\right)\right] = 1$;

π_{nix} is the probability that for child *n* the teacher chooses category x of ARS item *i*;

β_n is a person measure indicating the location of child *n* on the variable (e.g., Mathematical Thinking) being measured;

δ_i is the "difficulty" of ARS item *i*;

τ_k are response thresholds, or "step difficulties" for each response category on the rating scale;

m is the maximum category number,

x is the current category; and

j and *k* are suffixes that vary between 0 and *m*.

An easier to understand derivation of this model (Wright, 1999) is

$$\text{Log}(\pi_{nix}/\pi_{ni(x-1)}) = \beta_n - \delta_i - \tau_x \qquad (3.4)$$

β_n is comparable to the theta described in the chapter on the three-parameter IRT model used in estimating the scores for the direct measures.

3.2.1 Item Response Theory Estimation Using Winsteps

Winsteps software (Linacre and Wright, 2000), utilized to scale the Academic Rating Scale, uses joint maximum likelihood estimation. For initial estimates, the procedure PROX (approximation) is used. PROX assumes a normal distribution and does not take advantage of the ability of Rasch to calibrate measures independent of the sample characteristics(Wright & Masters, 1992). It provides a good starting point for the estimates. UCON (unconditional maximum likelihood) is used for the final iterations. UCON does not assume a normal distribution and performs a simultaneous estimation of the person and item parameters. With Winsteps, UCON is adjusted for the bias based on the length of the test (L/(L-1)) (Wright and Masters, 1982). Maximum scores are excluded for calibration of the items. Winsteps provides a variety of fit statistics and a factor analysis of the residuals.

Reliability estimates are provided for both the items and persons and indicate the replicability of the placement of the persons and items. The person reliability is analogous to Cronbach's alpha (table 6-1). Fit statistics are also provided for both persons and items (table 6-2). Both an information-weighted (infit) and an outlier sensitive (outfit) statistic are provided. The outfit mean square is sensitive to unexpected response on items far from the person's trait level. The infit mean square is weighted for the variance of the residual and thus is more influenced by unexpected responses close to the person's trait level (Linacre and Wright, 2000). The expected value for the mean square is 1.0. For samples larger than 1000, fit statistics greater than 1.1 indicate departures from expected response patterns that should be examined (Smith, Schumacker, and Bush, 1998).

Results of the IRT scaling of the teacher Academic Rating Scale are presented in chapter 6.

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 students) and a reference group (e.g., White students) when the two groups are "blocked" or matched on their total score. 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 items. 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 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 students with different patterns of exposure to content areas. For example, some students may be in schools where the kindergarten through third 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 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 ethnicity.

The DIF program was developed at ETS (Holland and Thayer, 1986) and was based on the Mantel-Haenszel odds-ratio (Mantel and Haenszel, 1959) and its associated chi-square. Basically, the Mantel-Haenszel (M-H) procedure forms odds-ratios from two-way frequency tables. 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 since there will be one table associated with each total number-right score from 0 to 20. Because of the two-stage, multi-form design of the ECLS-K tests, 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. The first dimension of each of the 41 tables is population subgroups (e.g., Whites vs. Blacks), and the remaining dimension is passing versus failing on a given item. Thus, the question that the M-H procedure addresses is whether or not members of the reference group (e.g., Whites), who have the same total ability estimate as members of the focal group (e.g., Blacks), have the same likelihood of passing the item in question. While 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. 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 M-H procedure provides a statistical test of whether or not the average odds-ratio significantly departs from unity for each item. If the probability is .05 or less, then one could say that there is statistical evidence for DIF on the item in question. The problem with this interpretation is two-fold. First, a very large number of statistical tests are being performed, one for each item for each pair of

subgroups, so low probabilities will be found occasionally even if no DIF is present. Second, if there are two relatively large samples involved, statistical significance will be virtually guaranteed.

Given these reservations, ETS has developed an "effect size" estimate that is not sample-size dependent. 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." Items labeled "A" or "B" either do not show statistically significant differential functioning for the two groups being compared or have differences that are too small to be important.

The fact that an item is identified by the DIF procedure does not mean that the item is necessarily unfair to any particular group. The DIF procedure is merely a statistical screening step that indicates 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 C-DIF items for evidence that the item may be measuring some extraneous dimension not consistent with the test framework. Items that attain C-level 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 third grade assessment items for six sets of contrast groups: males (reference group) compared with females (focal group), and White children (reference group) compared with five racial/ethnic minority groups: Black, Hispanic, Asian, Native American, and Multi-racial children. Statistics were computed for each item for which the minimum number of required responses, 200 observations for the smaller group, was available. The results of DIF analysis for the third 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 third grade round of testing. The types of scores derived from each of the assessments will be 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 four kindergarten and first grade rounds are reviewed, to the extent that they are relevant to interpretation of third 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 public-use data file. These analyses were carried out prior to final determination of cases eligible for the public-use file, and a few cases were deleted from the files. The psychometric results presented here are based on *all* children who had been tested at each round. Score statistics for all direct cognitive scores are presented in appendix A, with breakdowns by gender, race/ethnicity, socioeconomic status, and school type.

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 children's performance on the direct cognitive assessment are described in detail in this chapter. Number-right scores and IRT scale scores measure children'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 children'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 student who took the middle difficulty mathematics second-stage form would probably have gotten more questions correct if he or she had taken the easier low form and fewer if the more difficult high form had been administered. 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, either the kindergarten-first grade (K-1) or third grade version. The routing test in each subject area consisted of sets of items spanning a wide range of skills. For example, the reading routing test used for the four kindergarten and first grade rounds emphasized pre-reading skills, while the routing test in third grade contained easy and difficult decoding words, selecting the best word to complete a sentence, and a series of questions based on a reading passage. 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 fall-kindergarten through spring-first grade data collections, rounds 1 through 4, score comparisons *may* be made among these rounds. However, scores on the third grade routing tests were based on different and more difficult sets of items. The third grade routing test number-right scores should *not* be compared with the kindergarten or first grade routing test number-right scores.

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 children'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 students would have answered correctly at each point in time if they had taken all of the 154 questions in all of the first- and second-stage reading forms administered in all rounds, the 123 questions in all of the mathematics forms from all rounds, and the 62 third grade science items. 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 children'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 fall-kindergarten reading T-score of 45 represents a reading achievement level that is one-half of a standard deviation lower than the mean for the fall-kindergarten population represented by the assessed sample of ECLS-K participants. If the same child had a reading T-score of 50 in third grade, this would indicate that the child has made up his or her initial 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 public-use data files.

4.1.3 Item Cluster Scores

Several item cluster scores are reported for the reading and science assessments. These 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. The reading and science item cluster scores are described in sections 4.3.2 and 4.5.2.

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. Each cluster marked a learning milestone in reading or mathematics, agreed on by ECLS-K curriculum specialists. The sets of proficiency levels formed a hierarchical structure in the Piagetian sense in that the teaching sequence implied that one had to master the lower levels in the sequence before one could learn the material at the next higher level.

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 student 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 student 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 students in kindergarten through third 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 five rounds of data collection, less than 7 percent of reading response patterns and less than 5 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 eight reading and seven mathematics proficiency levels identified in the kindergarten through third grade assessments are described in sections 4.3.2 and 4.4.2, respectively. 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 or "I don't know" responses indicated lack of mastery. Questions that were answered with an explicit "I don't know" were treated as wrong, while omitted items were not counted. Since the ECLS-K direct cognitive child assessment was a two-stage design (where not all children were administered all items), and since more advanced assessment instruments were administered in third grade, 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 assessment set, and in part on different assessments being used for K-1 and third grade. 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 in that round of data collection on all of the items answered within the domain that included the incomplete cluster. 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. More than 80 percent of the "highest level" scores in both reading and mathematics were determined on the basis of item response data alone; the rest utilized IRT-based probabilities for some or all of the missing items. 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.

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, by imputation, or by a combination of methods. The highest proficiency level mastered implies that children demonstrated mastery of all lower levels and nonmastery of all higher levels. A zero score indicates non-mastery 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.

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 students' 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 level 1 probability at time 1 from the level 1 probability at time 2 indicates whether a student is advancing in mastery of the particular set of level 1 skills during this time interval. Thus, students' school experiences at selected times can be related to improvements in specific skills.

4.2 Motivation and Timing

An important issue in a low-stakes testing situation is motivation: whether the test results really represent the best efforts of the test takers. There are several pieces of evidence to support the conclusion that the ECLS-K participants were motivated to try their best. Field interviewers reported that children generally enjoyed the testing experience, took it seriously, and were cooperative. Another indication of motivation is the very small number of chance-level scores in the tables for the second-stage test forms. This suggests that children were putting effort into their responses rather than responding at random.

At the end of each testing session, assessors assigned a rating of each child's motivation, cooperation, and attention. Tables 4-1 to 4-3 show the distribution of these ratings in each round of data collection. These results show that assessors found the majority of children to be motivated, cooperative, and attentive during the sessions. At all rounds, nearly all children were perceived as cooperative (any of the highest three ratings). Motivation and attentiveness improved slightly from kindergarten to first to third grade, with over 90 percent of first and third graders rated in the highest three categories. Statistics in tables 4-1 to 4-3 include all children whose motivation, cooperation, and attention were rated by the assessors, even though not all received scores on the cognitive tests. Limited English proficiency, especially in the early rounds, was the primary reason for some children being excluded from the cognitive assessments.

There were no time limits on test sections; children were able to proceed at their own speed. Tests were discontinued only if children seemed unable or unwilling to continue. This approach resulted in scoreable tests for almost all of the children who started a testing session. Only about one-third of one percent of testing sessions could not be completed, primarily because of scheduling difficulties or children's mental or physical limitations. Of the completed assessments, nearly 95 percent were completed without special accommodations. The most common accommodation involved the scheduling/timing of the assessment, followed by assessment requirements in children's Individualized Education Plans (IEPs). More details on accommodations provided during data collection can be found in *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) User's Manual for the ECLS-K Third Grade Restricted-Use Data File and Electronic Code Book* (NCES 2003–003) and *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) User's Manual for the ECLS-K Public-Use Data File and Electronic Code Book* (NCES 2004–001). As the following tables

report, only a very small number of children who were assessed answered too few items for scores to be calculated.

Table 4-1. Child's overall motivation level during the assessment, in percent: Rounds 1 through 5: School years 1998–99, 1999–2000, and 2001–02

Category	Round 1	Round 2	Round 3	Round 4	Round 5
Number of cases	19,045	19,884	5,253	16,684	14,383
Very low: Child doesn't try or attempt many items, even with encouragement	1.7	1.6	1.0	1.2	1.4
Low: Child frequently says "I don't know" without even trying, consistent encouragement needed	9.9	10.4	7.5	8.1	6.8
Average: Child works on most items, says "I don't know" or refuses to answer items after s/he has begun doing some work or after making some attempt to figure the item out.	48.5	44.5	44.9	39.5	33.7
High: Child tries or attempts every item, including some of the most difficult.	29.8	30.7	32.5	31.4	35.3
Very High: Child tries or attempts every item, even the most difficult, appears interested in all the items, may need encouragement to move on to other items.	10.0	12.9	14.2	19.9	22.7
Very low + Low	11.6	11.9	8.5	9.3	8.3
Average + High + Very high	88.4	88.1	91.5	90.7	91.7

NOTE: Approximately 89 percent of the round 5 children were in third grade during the 2001–02 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. Percentages are unweighted. One child in round 1 received ratings for cooperation and attention but not for motivation. Details may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Fall 1998, spring 1999, fall 1999, spring 2000, and spring 2002.

Table 4-2. Child's overall cooperation during the assessment, in percent: Rounds 1 through 5: School years 1998–99, 1999–2000, and 2001–02

Category	Round 1	Round 2	Round 3	Round 4	Round 5
Number of cases	19,046	19,884	5,253	16,684	14,383
Very uncooperative: Child repeatedly refuses to comply.	1.1	0.6	0.4	0.8	0.8
Uncooperative: Child complies at least 50 percent of the time.	2.7	2.0	1.5	1.3	0.9
Matter of fact: Child complies at least 75 percent of the time.	22.7	23.5	22.1	23.2	14.5
Cooperative: Child complies with most (80-90 percent) requests and directives.	53.2	49.6	49.9	43.5	44.1
Very cooperative: Child complies with all requests and directives in first request.	20.3	24.3	26.1	31.1	39.7
Very uncooperative + Uncooperative	3.8	2.6	1.9	2.2	1.7
Matter of fact + Cooperative + Very cooperative	96.2	97.4	98.1	97.8	98.3

NOTE: Approximately 89 percent of the round 5 children were in third grade during the 2001–02 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. Percentages are unweighted. Details may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, and spring 2002.

Table 4-3. Child's overall attention level during the assessment, in percent: Rounds 1 through 5: School years 1998–99, 1999–2000, and 2001–02

Category	Round 1	Round 2	Round 3	Round 4	Round 5
Number of cases	19,046	19,884	5,253	16,684	14,383
Unable to attend: Child needs ongoing redirection to the task.	0.6	0.6	0.3	0.3	0.4
Difficulty attending: Child is distracted easily and often requires redirection.	13.6	11.4	8.0	9.4	9.0
Attentive: Child attends the majority of the time, when distracted child returns to task with redirection.	43.3	37.9	37.9	35.7	31.5
Very attentive: Child may momentarily be distracted but is able to return to the task on his/her own.	31.0	33.9	35.2	32.1	33.6
Complete and full attention: Child is able to ignore any distractions.	11.5	16.3	18.7	22.5	25.5
Unable to attend + Difficulty attending	14.2	12.0	8.3	9.7	9.4
Attentive + Very attentive + Complete and full attention	85.8	88.0	91.7	90.3	90.6

NOTE: Approximately 89 percent of the round 5 children were in third grade during the 2001–02 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. Percentages are unweighted. Details may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, and spring 2002.

4.3 Reading Assessment

The third grade reading test emphasized reading comprehension, with the majority of questions based on one of several reading passages. Additional questions tapped basic skills, including decoding and vocabulary. Children began the reading assessment with a routing test of 15 items, 5 of which were based on a short reading selection. The score on the routing test was used to select one of three second-stage forms, of varying difficulty, each consisting of 4 (low form) or 5 (middle and high forms) reading passages with associated questions, plus 5 or 6 individual decoding vocabulary items.

4.3.1 Samples and Operating Characteristics

Table 4-4 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 1–4, fall-kindergarten through spring-first grade. A new set of assessment forms suitable for third graders was used in round 5. The small sample size reported at round 3 in table 4-4 reflects the fact that only a subsample of the fall-first grade longitudinal cohort was assessed at this point in time. 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. Scores were calculated only for children who attempted at least 10 items in the routing test and second-stage form combined. Children who lacked sufficient English proficiency to pass the English language screening test, administered in rounds 1 through 4, were excluded from the reading assessment.

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 students took the highest level of the second-stage forms. The third grade assessment developed for round 5 was 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.

Table 4-4. Reading assessment: Samples and operating characteristics: Rounds 1 through 5: School years 1998–99, 1999–2000, and 2001–02

Characteristics	Round 1	Round 2	Round 3	Round 4	Round 5
Total	17,630	18,944	5,054	16,340	14,286
Too few items	44	19	0	2	134
Number taking low form	13,355 (76%)	6,521 (34%)	1,062 (21%)	618 (4%)	3,540 (25%)
Number taking middle form	3,620 (21%)	8,906 (47%)	2,334 (46%)	2,371 (15%)	8,032 (56%)
Number taking high form	654 (4%)	3,517 (19%)	1,657 (33%)	13,351 (82%)	2,714 (19%)
Percent perfect score routing test	.3	1.7	4.9	23.6	3.4
Percent perfect score low form	0.0	0.1	0.4	1.6	0.0
Percent perfect score middle form	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
Percent less than chance routing test	22.6	3.7	2.1	0.3	0.4
Percent less than chance low form	0.9	0.5	0.2	0.6	3.6
Percent less than chance middle form	0.5	0.3	0.1	0.1	0.2
Percent less than chance high form	0.5	1.7	2.3	0.4	0.0

NOTE: Rounds 1–4 used the same set of assessment forms; Round 5 forms were a different set developed for third grade. Approximately 89 percent of the round 5 children were in third grade during the 2001–02 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. "Too few items" refers to the number of children who did not attempt a sufficient number of reading items to generate a reliable score. Percentages are unweighted. Form counts many not sum to total because a few children answered enough items in the routing test to receive a reading score, but no items in a second-stage form.

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, and spring 2002.

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 in table 4-4 demonstrate that the two-stage test design accomplished its objective of avoiding floor and ceiling effects. The percentages of perfect scores were all close to zero with exception of the round 4 routing test. Although about 23 percent of children had perfect scores on the routing test in round 4, 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 combined. Since there was no ceiling effect problem in the high-level second-stage form (virtually no perfect scores in any round), the perfect routing test scores did not have the potential to create a ceiling effect. Table 4-4 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. Again, their final scores reflected performance on the combined set of routing and second-stage items. A small floor effect 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 4 correct answers on the routing and second-stage forms combined.

4.3.2 Scores Unique to the Reading Assessment: Cluster Scores and Proficiency Levels

Cluster scores. 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 (zero to three) for the three items. The print familiarity cluster score is documented in *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Psychometric Report for Kindergarten Through the First Grade* (NCES 2002–05) and is included in the K-1 public-use data files (*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 Files and Electronic Code Book,* NCES 2002–148). These items were not included in the third grade reading forms because nearly all children had mastered them by the end of first grade.

A set of four relatively difficult decoding items is reported for the third grade assessment. These were words that were unlikely to be in most children's everyday vocabulary, but could be sounded out phonetically.

Proficiency levels. The following eight 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; and

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.

The test items on which levels 1–3 were based appeared only in the K-1 assessments, not in third grade, while levels 6–8 were defined based on third grade assessment items. Level 4 and 5 items were included in both the K-1 and third grade assessment forms. IRT procedures described in sections 3.1 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 4-5 presents reliability statistics for the third grade reading assessment. K-1 reliabilities are included in the table for comparison purposes. 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.

Table 4-5. Reading assessment reliabilities, rounds 1 through 5: School years 1998–99, 1999–2000, and 2001–02

Reliability Measure	Round 1	Round 2	Round 3	Round 4	Round 5
Alpha routing	.86	.88	.88	.86	.75
Alpha low form	.69	.69	.71	.72	.83
Alpha middle form	.70	.72	.74	.78	.84
Alpha high form	.90	.88	.93	.92	.79
Split-half: Decoding score	†	†	†	†	.67
Split-half: Proficiency level 1	.83	.79	.77	.78	†
Split-half: Proficiency level 2	.76	.76	.73	.70	†
Split-half: Proficiency level 3	.72	.76	.76	.68	†
Split-half: Proficiency level 4	.78	.77	.80	.78	.56
Split-half: Proficiency level 5	.60	.69	.73	.73	.66
Split-half: Proficiency level 6	†	†	†	†	.48
Split-half: Proficiency level 7	†	†	†	†	.48
Split-half: Proficiency level 8	†	†	†	†	.63
Reliability of theta	.93	.95	.96	.97	.94
Percent agreement of highest proficiency level mastered:					
Percent exact agreement	68	57	57	59	54
Percent exact + off by 1	97	95	95	95	94

† Not applicable.
NOTE: Statistics are unweighted. Approximately 89 percent of the round 5 children were in third grade during the 2001–2002 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth 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, and spring 2002.

Internal consistency coefficients for third grade are comparable to those obtained for K-1. The third grade alpha coefficient for the routing test is somewhat lower than that of the earlier rounds, at least in part due to the third grade form having fewer items (15) than the 20 items in the K-1 version. The alpha coefficients for the K-1 second-stage forms are generally lower than those of the routing test due to the restriction in range 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 second-stage form, which had much greater variance than did the other forms, did the K-1 alpha coefficients approach or exceed .90. This tendency for the K-1 second-stage forms to have lower alphas due to restriction in range was counteracted in third grade by the greater number of items in the third grade second-stage forms, in comparison with 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 cluster was present only in the third grade assessment, not in the earlier rounds. Split-half reliabilities are presented for the individual proficiency level scores for information only since "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 6–8 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.

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. 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.

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 student demonstrating mastery of the first five reading levels but not the remaining three 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 student's highest level mastered could be determined on the basis of actual item response data alone for more than 80 percent of the 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 students. The percent 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.

4.3.4 Score Statistics

Table 4-6 presents 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 154 test items. Four additional items were deleted from scoring because they turned out to be much too difficult for the third graders and it was not possible to estimate stable item parameters for them. The IRT procedures described earlier allowed these estimates to be computed based on the subset of questions actually administered to each child at each round. 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. The variability in reading performance found at the end of first and third grades is greater than that observed in the earlier rounds. Score statistics for all reading scores, with breakdowns by population subgroups, are presented in appendix A.

Table 4-6. Reading assessment scale score means and standard deviations, rounds 1 through 5: School years 1998–99, 1999–2000, and 2001–02

	Round 1	Round 2	Round 3	Round 4	Round 5
Scale score mean	26.9	37.9	44.3	66.6	106.1
Scale score standard deviation	9.7	13.0	16.5	20.8	20.7

NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3A5W0, C4A3W0, C5CW0). Approximately 89 percent of the round 5 children were in third grade during the 2001–02 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. Estimates for kindergarten through third grade have been put on a common scale to support comparisons. The range of values: 0–154.
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, and spring 2002.

4.3.5 Differential Item Functioning

Section 3.3 explains the DIF procedures used for identifying test items that perform differentially for population subgroups. Table 4-7 summarizes the results of the DIF analysis of the third grade reading items. The largest number of C-DIF[1] items was found for performance comparisons of White vs. Asian children, with some items favoring the focal group (Asian children) and some the reference group (White children). There are several reasons for these numbers to be larger than those for the other subgroup contrasts. First, the field test of third grade items had too few Asian participants for DIF analysis to be carried out on field test data, so that items with the potential for White/Asian DIF were not identified and removed from consideration for the third grade assessments. Second, many of the Asian children came from a language minority background. The five items on which Asian children performed relatively better than expected were basic skills items (decoding and spelling). The three questions that were relatively harder for Asian children involved inferences based on stories. (Compare these numbers with the small number of C-DIF items, favoring either the focal group or the reference group, for Asian children in the mathematics and science assessments described below.) There were insufficient numbers of Native American and multiracial children in the sample for DIF statistics to be computed for test items that appeared in only one second-stage form.

Table 4-7. Reading assessment: Differential item functioning, third grade: School year 2001–02

Reference group: Focal group:	Male Female	White Black	White Hispanic	White Asian	White Native American	White Multi-racial
Number of C-DIF[1] items favoring focal group	1	0	2	5	0	0
Number of C-DIF items favoring reference group	2	0	1	3	0	0

[1] DIF having an effect size of 1.5 or greater.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

[1] ETS has developed an "effect size" estimate that is not sample-size dependent. 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."

It should be kept in mind that there were 90 reading items in the third grade reading assessment forms, and six sets of comparison groups. Even with insufficient sample sizes for some of the DIF statistics to be computed for some groups, several hundred comparisons were made. 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.

All C-DIF reading items were reviewed and found to be relevant to the construct being measured by the assessment, so all were retained in the scoring procedures.

4.4 Mathematics Assessment

The third grade mathematics framework specifications were quite similar to those for kindergarten and first grade, in terms of percentages of items in each content strand for the whole item pool. The easier items in the routing test and low second-stage form tended to focus on number sense, properties, and operations, while the more difficult forms contained a larger proportion of measurement, algebra, and geometry items. Greater emphasis was placed on problem solving in third grade compared with K-1. Children began the mathematics assessment with a routing test of 17 items. The score on the routing test was used to select one of three second-stage forms, of varying difficulty, each consisting of 25 (low form) or 24 (middle and high forms) items.

4.4.1 Samples and Operating Characteristics

Table 4-8 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 1-4, 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.

A more advanced set of assessment forms, entirely in English, was developed for third grade. Scores were calculated only for children who attempted at least ten items in the routing test and second-stage form combined.

The third grade assessment developed for round 5 was 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. Fewer third graders were routed to the middle difficulty second-stage form than anticipated, and more to the low and high forms. This discrepancy may be due to greater variability in the emphasis placed on mathematics skills (compared with reading) by different schools in the early elementary years. Again, the important point here is not matching the anticipated 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 very low percentages of perfect and below-chance scores observed in the assessments demonstrate that this strategy was successful in avoiding floor and ceiling effects.

Table 4-8. Mathematics assessment: samples and operating characteristics, rounds 1 through 5: School years 1998–99, 1999-2000, and 2001–02

Characteristics	Round 1	Round 2	Round 3	Round 4	Round 5
Total	18,641	19,657	5,226	16,647	14,380
Too few items	21	15	0	2	29
Number taking low form	14,380 (77%)	8,444 (43%)	1,353 (26%)	1,097 (7%)	4,229 (29%)
Number taking middle form	3,123 (17%)	6,169 (31%)	1,521 (29%)	2,317 (14%)	5,344 (37%)
Number taking high form	1,136 (6%)	5,042 (26%)	2,351 (45%)	13,233 (79%)	4,804 (33%)
Percent perfect score routing test	0.1	0.4	1.5	7.9	1.6
Percent perfect score low form	0.1	0.4	1.0	2.5	0.0
Percent perfect score middle form	0.0	0.0	0.0	0.3	0.1
Percent perfect score high form	0.0	0.0	0.0	0.1	0.0
Percent less than chance routing test	15.3	3.1	1.6	0.3	1.3
Percent less than chance low form	0.9	0.3	0.1	0.3	0.3
Percent less than chance middle form	0.1	0.0	0.0	0.0	0.1
Percent less than chance high form	0.1	0.0	0.0	0.0	0.1

NOTE: Rounds 1–4 used the same set of assessment forms; Round 5 forms were a different set developed for third grade. Some children in rounds 1–4 received a Spanish translation of the mathematics assessment; in round 5, all assessments were in English. Approximately 89 percent of the round 5 children were in third grade during the 2001–02 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. "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. Form counts may not sum to totals because a few children answered enough items in the routing test to receive a test score, but no items in a second-stage form.

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, and spring 2002.

4.4.2 Scores Unique to the Mathematics Assessment: Proficiency Levels

The following seven 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; and

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

The test items on which levels 1–3 were based appear only in the K-1 assessments, not in third grade, while levels 6 and 7 are defined based on third grade assessment items. Level 4 and 5 items were included in both the K-1 and third grade assessment forms. IRT procedures were used to obtain probability estimates for all levels at all rounds, as described in sections 3.1 and 5.2, so that longitudinal gains in specific skills could be measured.

4.4.3 Reliabilities

Table 4-9 presents reliability statistics for the third grade mathematics assessment. K-1 reliabilities are included in the table for comparison purposes.

All other things being equal (e.g., type and difficulty of test items), internal consistency coefficiencts tend to be higher for longer tests, and lower when the ability range of the test takers is

Table 4-9. Mathematics assessment reliabilities, rounds 1 through 5: School years 1998–99, 1999–2000, and 2001–02

Reliability measure	Round 1	Round 2	Round 3	Round 4	Round 5
Alpha routing	.78	.81	.83	.80	.86
Alpha low form	.70	.66	.66	.71	.77
Alpha middle form	.66	.67	.66	.66	.72
Alpha high form	.80	.80	.83	.82	.73
Split-half: Proficiency level 1	.41	.27	.26	.26	†
Split-half: Proficiency level 2	.58	.49	.51	.32	†
Split-half: Proficiency level 3	.63	.66	.67	.59	†
Split-half: Proficiency level 4	.54	.63	.66	.63	.43
Split-half: Proficiency level 5	.46	.53	.61	.65	.67
Split-half: Proficiency level 6	†	†	†	†	†
Split-half: Proficiency level 7	†	†	†	†	.43
Reliability of theta	.92	.94	.94	.94	.95
Percent agreement of highest proficiency level mastered:					
Percent exact agreement	58	56	56	61	61
Percent exact + off by 1	97	96	96	98	98

† Not applicable.
NOTE: Statistics are unweighted. Approximately 89 percent of the round 5 children were in third grade during the 2001–2002 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. The four test items for mathematics proficiency level 6 did not all appear in the same test form, so no complete data cases were available for evaluation of split half reliability.
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, and spring 2002.

restricted. The internal consistency (alpha) coefficient for the third grade mathematics routing test was slightly higher than that of earlier forms, probably partly due to a slightly longer test in third grade (17 items vs. 16 items in K-1), and partly because of greater variability in the mathematics achievement of third graders compared with earlier rounds. The third grade second-stage mathematics forms have lower alpha coefficients than the routing test because of the restricted variance 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 grade second-stage forms all had about the same number of items, 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: level 1–3 items were present only in the K-1 mathematics assessment, while level 6 and 7 items appeared only in

the third grade forms. Items for levels 4 and 5 were included in the assessments for all five rounds. There is no split-half reliability presented for proficiency level 6 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 primarily because the mathematics items were not as homogeneous with respect to similarity of content and skill demand as was the case with reading. In third grade, the items for mathematics proficiency level 5 were included in the routing test, which was taken by all children, and the reliability was similar to that of the earlier rounds. Level 4 items appeared only in the low second-stage form in third grade, and the restriction in range of ability resulted in a lower split-half reliability for this cluster. Similarly, only high second-stage form test takers in third grade took all of the level 7 items, resulting in a relatively low split-half reliability for this level as well.

The percentages of agreement between methods in determining the highest mathematics proficiency level mastered were slightly higher than those for reading, both for percentage of exact agreement, and percentage of agreement within one level. See section 4.3.3 for a detailed explanation of how this score was computed and evaluated.

Also similar to the reading test, the reliabilities of the theta scores were in the mid 90s. The reliability of theta applies to the scale scores and proficiency probabilities as well, since these scores are nonlinear transformations of the thetas.

4.4.4 Score Statistics

The scale score means presented in table 4-10 represent estimates of the number of correct answers that would have been expected if each child had been given all of the 123 mathematics items in the pool, that is, all items that appeared in any of the K-1 and/or third grade test forms and were scored. 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. Score statistics for the mathematics scores and breakdowns by population subgroups are presented in appendix A.

Table 4-10. Mathematics assessment scale score means and standard deviations, rounds 1 through 5: School years 1998–99, 1999–2000, and 2001–02

Item	Round 1	Round 2	Round 3	Round 4	Round 5
Scale score mean	21.0	30.8	37.5	53.7	83.2
Scale score standard deviation	8.7	11.3	13.4	16.1	18.3

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Approximately 89 percent of the round 5 children were in third grade during the 2001–02 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. Estimates for kindergarten through third grade have been put on a common scale to support comparisons. The range of values: 0–123.
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, and spring 2002.

In general, geometry items tended to have lower *r*-biserials and IRT "a" parameters than the other item categories. Two of the weakest items were deleted from the third grade scoring procedures because they contributed very little in terms of measurement objectives. Both items were more closely related to spatial ability than to curriculum topics. These items had also performed relatively poorly in the third grade field test, but were selected for the operational forms because of a shortage of suitable geometry items that were needed to meet target numbers specified in the mathematics framework. Because these two items failed to meet psychometric standards for third grade scoring, they were deleted from the pool.

Two additional items administered in third grade were deleted from the scores because of DIF findings, as described in the next section.

4.4.5 Differential Item Functioning

Table 4-11 presents counts of the C-DIF items for the third grade mathematics forms. One of the two items with DIF favoring the reference group, White children, in the White/Asian contrast was the same item identified as having DIF favoring White children compared with Black and Hispanic children. This item, which was used in the third grade low second-stage form, had a relatively high language load and may possibly have been confusing to children with limited verbal ability. Although it had a high *r*-biserial, indicating a strong relationship with the rest of the mathematics items in the form, it was deleted from scoring. The other item found to have C-level DIF against Asian children was also in the low second-stage form. This item, a geometry item that relied on spatial ability, was judged to not be strongly curriculum related, and was deleted from scoring.

Table 4-11. Mathematics assessment: Differential item functioning, third grade: School year 2001–02

Reference group: Focal group:	Male Female	White Black	White Hispanic	White Asian	White Native American	White multi-racial
Number of C-DIF[1] items favoring focal group	1	0	0	1	0	0
Number of C-DIF items favoring reference group	0	1	1	2	0	0

[1] DIF having an effect size of 1.5 or greater.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

There were insufficient numbers of Native American and multiracial children in the sample for DIF statistics to be computed for about half of the mathematics items, which appeared in only one second-stage form. See section 3.3 for an explanation of DIF procedures.

4.5 Science Assessment

The third grade science assessment consisted of a 15-item routing test followed by three second-stage forms of 20 items each. Content of the science questions was approximately equally divided among life science, earth science, and physical science strands. The science assessment was added to the ECLS-K cognitive battery for third grade; thus there is no longitudinal score scale and no comparisons to be made with K-1 results. When the next round of data is collected in fifth grade, a longitudinal score scale will be developed.

4.5.1 Samples and Operating Characteristics

Table 4-12 presents sample counts and operating characteristics of the third grade science forms. Scores were calculated only for children who attempted at least ten items.

Slightly more children were routed to the low second-stage form, and slightly fewer 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 absence of floor and ceiling

Table 4-12. Science assessment: Samples and operating characteristics, round 5: School year 2001–02

Characteristics	Round 5
Total	14,357
Too few items	41
Number taking low form	4,199 (29%)
Number taking middle form	7,204 (50%)
Number taking high form	2,952 (21%)
Percent perfect score routing test	1.5
Percent perfect score low form	0.3
Percent perfect score middle form	0.0
Percent perfect score high form	0.0
Percent less than chance routing test	4.7
Percent less than chance low form	1.7
Percent less than chance middle form	0.6
Percent less than chance high form	0.8

NOTE: No science assessment was conducted in rounds 1–4. The round 5 assessment was developed for third grade. Percentages are unweighted. Approximately 89 percent of the round 5 children were in third grade during the 2001–2002 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. "Too few items" refers to the number of children who did not attempt a sufficient number of reading items to generate a reliable score. Form counts many not sum to totals because a few children answered enough items in the routing test to receive a test score, but no items in a second-stage form.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

effects. Although 4.7 percent of children had below-chance scores on the routing test, and 1.7 percent on the low second-stage form, less than half of one percent of the low form children had below-chance scores for the two sections combined. No children received perfect scores for the total assessment.

4.5.2 Scores Unique to the Science Assessment: Cluster Scores

The science assessment does not have sets of proficiency levels in the same sense as the hierarchical levels for reading and mathematics. Different states and different schools may have quite different sequences for teaching science units. Many science topics are independent of each other, so there is no logical interpretation that would imply that mastery of a set of questions would imply mastery of a set based on different topics.

The 15 routing form items of the third grade science assessment tapped a range of basic concepts, with 5 questions each in life science, physical science, and earth science:

- **Life Science:** a sample of concepts related to anatomy/health, animal characteristics/ behavior, and ecology;
- **Physical Science:** a sample of concepts related to states of matter, sound, physical characteristics, and the scientific method; and
- **Earth Science:** a sample of concepts related to the solar system, earth, soil, minerals, and weather.

Scores consisting of simple counts of number right for the 5 items were computed for each of the three clusters. Children who omitted more than 2 items in a cluster were not scored. The items were not selected to have comparable levels of difficulty within each set. For example, the mean of 3.0 for the life science cluster compared with 2.6 for earth 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 earth science. With only 5 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. Factor analysis of the science routing test items did not result in identification of any underlying factors related to the life science/physical science/earth science categories.

4.5.3 Reliabilities

Table 4-13 presents reliability statistics for the third grade science assessment. Alpha coefficients for the routing test and second-stage forms are somewhat lower than those for reading and mathematics because the science assessment consisted of more diverse content, and had fewer items in the second-stage forms. As in reading and mathematics, the second-stage alpha coefficients were depressed in comparison with the routing test because the range of ability within each form was restricted. The children taking each of these forms are a more homogeneous group with respect to science performance, so the score variance, and thus the alpha coefficient, are lower than they would have been if the whole sample of children had taken each form. Scores for the second-stage forms are not reported separately.

Table 4-13. Science assessment reliabilities, third grade: School year 2001–02

Reliability measure	Round 5
Alpha routing	.75
Alpha low form	.70
Alpha middle form	.61
Alpha high form	.60
Split-half: Life Science	.59
Split-half: Physical Science	.49
Split-half: Earth Science	.46
Reliability of theta	.88

NOTE: Statistics are unweighted. Approximately 89 percent of the round 5 children were in third grade during the 2001–02 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth 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.

Diversity of content accounted for the relatively low split-half reliability of the science clusters compared with the decoding score in the reading assessment. 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 for the same reason.

4.5.4 Score Statistics

Third grade science scale score statistics are presented in table 4-14 and represent the number of correct answers that would have been expected if each child had been given all of the 62 items in all of the test forms. Despite the diversity of content in the assessment, all items had acceptable fit to the IRT model. Score statistics for all science scores and breakdowns by population subgroups are presented in appendix A.

Table 4-14. Science scale score mean and standard deviation, third grade: School year 2001–02

Item	Round 5
Scale score mean	33.5
Scale score standard deviation	10.0

NOTE: Table estimates are based on round 5 cross-sectional weight (C5CW0). Approximately 89 percent of the round 5 children were in third grade during the 2001–02 school year, 9 percent were in second grade, and fewer than 1 percent were in fourth grade. The range of values is 0–62.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

4.5.5 Differential Item Functioning

Table 4-15 summarizes the results of the DIF analysis of the third grade science items. Only three items were identified as having C-DIF, and two of the three favored the focal group. Of the two items relating to the White/Native American contrast, one favored the reference group and one the focal group. This finding may be a consequence of some instability due to the small numbers of Native American children in the sample (247), only slightly exceeding the minimum of 200 required for DIF computation. None of the three C-DIF items had a DIF category other than "A" for any other contrast. There were insufficient numbers of Native American and multiracial children in the sample for DIF statistics to be computed for test items that appeared in only one second-stage form. All C-DIF science items were reviewed and found to be relevant to the construct being measured by the assessment, so all were retained in the scoring procedures.

Table 4-15. Science assessment: Differential item functioning, third grade: School year 2001–02

Reference group: Focal group:	Male Female	White Black	White Hispanic	White Asian	White Native American	White Multi- Racial
Number of C-DIF[1] items favoring focal group	0	0	0	1	1	0
Number of C-DIF items favoring reference group	0	0	0	0	1	0

[1] DIF having an effect size of 1.5 or greater.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

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

4.6 Intercorrelations among the Direct Cognitive Measures

Evidence for the construct validity of the direct measures of children's achievement can be generated by observing certain consistent correlational patterns within and across the rounds. The correlation between the third grade reading and mathematics scores was .73, which is consistent with the correlations found in the kindergarten and first grade rounds and continues the slight decline in correlations during that time (.77 to .74). The science assessment, new in third grade, correlated .72 with

both reading and mathematics. This is a somewhat stronger relationship than had been found for the kindergarten and first grade general knowledge test (.57 to .59 with reading; .64 to .67 with mathematics). The higher correlation is probably due to the third grade science assessment being more strongly curriculum-related than the K-1 general knowledge test. The general knowledge test consisted of a mix of science and social studies concepts, many of which most children could have encountered outside of school.

See chapter 6 for a discussion of the discriminant and convergent validity of the direct and indirect cognitive measures.

This page is intentionally left blank.

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-third grade. (Science assessments were not conducted prior to third grade.) The potential impact of the absence of a second grade data collection round was discussed in chapter 2, and is summarized below. This chapter will describe the collection of reading and mathematics data for a small sample of second graders, and show how the bridge sample data were used to supplement the longitudinal sample data in establishing vertical scales for measurement of gain. The development of the longitudinal scales, including analysis of common items, will be described. The final section of the chapter will focus on applications: choosing the appropriate scores for analysis and interpreting gain statistics.

5.1 Bridge Study

Chapter 2, section 2.1.5, documents the gap in ability levels that was anticipated due to the absence of the second grade data collection from the longitudinal design due to budgetary constraints. 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 need not be part of the ECLS-K longitudinal cohort. They need only have ability levels in the range where the ECLS-K longitudinal sample data might be sparse, and take sets of cognitive test items that include the items designed to link the first and third grade rounds. As described in chapter 2, field test data collected for second and third graders demonstrated that the longitudinal sample third grade assessment forms would be appropriate for the second graders in the bridge sample.

In order to bridge the expected ability gap, reading and mathematics assessments were administered to a sample of approximately 900 second graders in 43 schools. While the bridge sample was a convenience sample and was not designed to be nationally representative, efforts were made to

include a diverse sample of children and schools. About 77 percent of the bridge sample children were White and 23 percent minority; 30 public schools and 13 religious or other private schools participated; and attention was given to recruiting schools spanning a wide range of socioeconomic (SES) levels. However, because the bridge sample participants did not constitute a nationally representative sample of second graders and were not part of the ECLS-K longitudinal sample, their assessment scores are not included in released data files. 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, in order that reliable gain scores could be estimated for the first-to-third graders in the ECLS-K sample.

The results of the bridge study reading and mathematics assessments are summarized below. Table 5-1 presents operating characteristics for the bridge sample second graders. The line labeled "Too few items" refers to the number of children who did not attempt a sufficient number of reading and mathematics items to generate a reliable score. Scores were calculated only for children who attempted at least ten items in the routing test and second-stage form combined.

Table 5-1. Bridge sample operating characteristics: School year 2001–02

Characteristics	Reading	Mathematics
Total	904	902
Too few items	3	5
Number taking low form	331 (37%)	486 (54%)
Number taking middle form	512 (57%)	308 (34%)
Number taking high form	61 (7%)	108 (12%)
Percent perfect score routing test	0.8	0.0
Percent perfect score low form	0.0	0.0
Percent perfect score middle form	0.0	0.0
Percent perfect score high form	0.0	0.0
Percent less than chance routing test	0.2	2.2
Percent less than chance low form	1.5	0.2
Percent less than chance middle form	0.8	0.0
Percent less than chance high form	0.0	0.0

NOTE: Bridge sample assessments used the ECLS-K third grade assessment forms. Percentages are unweighted.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), second grade bridge study, spring 2002.

It had been anticipated that the majority of second graders would be routed to the low second-stage forms of each assessment. This turned out to be true for the mathematics assessment, by a small majority, but not for the reading assessment. The 57 percent of bridge sample second graders who received the middle level reading form was about the same as for the longitudinal sample third graders (56 percent), although fewer third graders were routed to the low form (25 percent) and more to the high form (19 percent). Scale score statistics (presented in the next section) showed the average bridge sample achievement levels, in both reading and mathematics, to be closer to the results of the ECLS-K third graders than to those of the spring-first grade sample. This was especially true for reading scores. This is not surprising, for two reasons. First, it is consistent with the idea that reading is heavily emphasized in the early elementary years. As a result, many children read fluently by the end of second grade, and progress in reading may slow down in the later grades as emphasis on mathematics and other subject areas increases. Second, efforts were made to represent all levels of ability in the bridge sample. The inclusion of several high SES schools in the sample to satisfy this requirement may have resulted in achievement levels that were, on average, somewhat higher than might be found in a nationally representative sample of second graders.

The overlap of items between the low and middle second stage forms and the discrete nature of the routing test cut points means that for children near the cut points, the selection of one or the other form is not critical. Table 5-1 shows that there were no ceiling or floor effect problems for participants in the bridge study. Table 5-2 shows that for children routed to the middle and high forms in both reading and mathematics, the average number correct was lower for second than for third graders taking the same form. In other words, comparisons of percentages routing to particular forms tell only part of the story: there is still a wide range of ability being measured *within* each form, and differences between second and third grade children taking the same assessment forms are apparent. For children routed to the low second-stage form, average number correct was similar for second and third graders, but a much larger proportion of second graders than third graders received the low form (see tables 4-4 and 4-8 for comparisons).

Table 5-2. Average number correct for second and third graders on comparable test sections: School year 2001–02

Test section	Reading		Mathematics	
	Second grade bridge	Third grade longitudinal	Second grade bridge	Third grade longitudinal
Routing test	9.1	10.1	6.5	9.1
Low form	9.4 (37%)	9.5 (25%)	14.6 (54%)	13.9 (29%)
Middle form	20.4 (57%)	22.6 (56%)	12.4 (34%)	13.6 (37%)
High form	21.5 (7%)	24 (19%)	8.7 (12%)	11.5 (33%)

NOTE: All items appeared in routing tests in both K-1 and third grade assessments. Percentages are unweighted.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), second grade bridge study and spring 2002.

Table 5-3 shows percent correct on four reading and four mathematics items administered to all spring-first graders, bridge sample second graders, and third graders. These items were selected from a larger pool of items common to the K-1 and third grade assessment forms because all of these items appeared in the routing sections in both versions. As a result, they were taken by all children who participated in the assessments, and percentages may be compared. It is not meaningful to compare percent correct for items that were common to both assessments but appeared in second-stage forms and were taken by subsets of children selected according to performance. (Two additional reading items were common to the K-1 and third grade routing sections, but they were too easy for both second and third graders to show useful comparisons.)

Table 5-3. Average percent correct on items common to K-1 and third grade assessments, spring-first grade and spring-third grade: School years 1999–2000 and 2001–02

Percent correct	Reading			Mathematics		
	Spring-first grade	Second grade bridge	Spring-third grade	Spring-first grade	Second grade bridge	Spring -third grade
Common item #1	68	90	95	50	74	81
Common item #2	60	93	92	39	68	82
Common item #3	64	92	94	26	53	72
Common item #4	56	90	93	35	62	78

NOTE: All items appeared in routing tests in both K-1 and third grade assessments. Percentages 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 2000, second grade bridge and spring 2002.

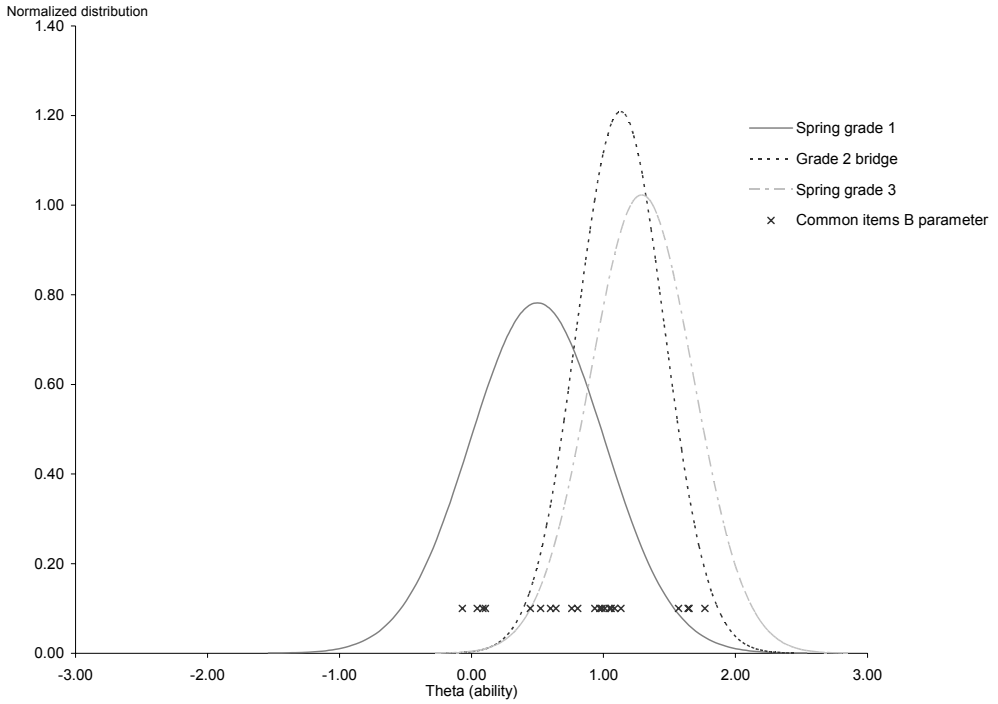
The percentages in table 5-3 show substantial differences between the longitudinal sample first graders and the bridge sample second graders for the four common reading items, but only small

differences between second and third graders. The mathematics percentages, by contrast, show sizeable differences in both the first/second and second/third grade contrasts. Ideally, comparisons of gaps would be most meaningful for items with first grade percentages that are similar for reading and mathematics, but no such items existed that were taken by the whole sample of children on each occasion. However, items representing a range of easier and harder difficulty were present in the tests administered to the children, so estimates did not need to be based solely on the common items shown in the table.

The bridge sample results suggest that the children tested may have been, on average, somewhat higher achievers than would be found in a nationally representative sample of second graders. Their mean ability levels are closer to those of the ECLS-K third graders than first graders. This conclusion appears to be more apparent for the reading than for the mathematics tests. Two caveats previously mentioned bear repeating. First, any differences in relative curriculum emphasis in first through third grades would affect this conclusion, but are confounded with the test results and cannot be evaluated independently. Second, differences in performance on common items are confounded with routing patterns, resulting in item statistics that are based on dissimilar subsets of children on each occasion.

Despite these limitations in evaluating the bridge sample as an independent data point, it is clear that the data are appropriate for the intended purpose of supplementing the data required for establishing the longitudinal scale linking first and third grade. Figures 5-1 and 5-2 illustrate the distributions of ability levels for the longitudinal sample spring first graders, the bridge sample children, and the spring-third graders. For both reading and mathematics, the ability levels of the longitudinal sample overlap for about the highest 20 percent of first graders and the lowest 20 percent of third graders. It is clear from the graphs that the bridge sample children represent levels of ability that lie between the preponderance of first and third grade scores. While they do not appear to lie *half-way* between the longitudinal sample grades, as noted above, there is no reason to expect them to do so. The symbols in the diagrams show the locations of the difficulty parameters of the common items that anchor the longitudinal scales. Note that the difficulty parameters of the 13 common items in mathematics do not span as much of the second grade ability range as do the 22 common reading items. Nine of the 13 mathematics items do, however, fall in the range of abilities where the bridge sample was designed to supplement the relatively sparse data in the tails of the first and third grade distributions.

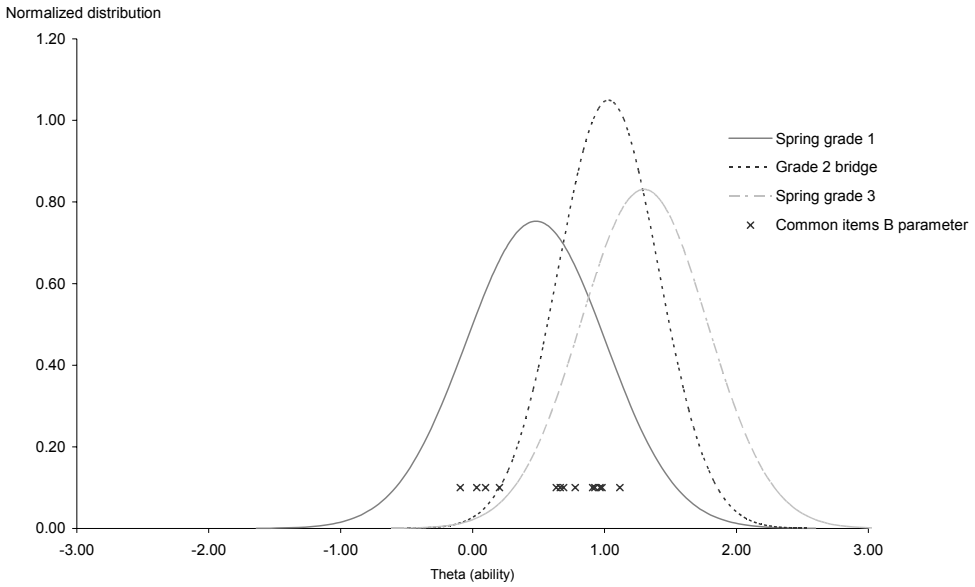
Figure 5-1. Normal distributions of ability for adjacent samples, and difficulty parameters of common items: Reading (first grade, second grade bridge, and third grade): School years 1999–2000 and 2001–02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2000, second grade bridge study and spring 2002.

Reliability and DIF statistics were not evaluated for the bridge sample, and score statistics are not presented because the bridge data were used for stabilizing item parameters rather than for evaluating children's achievement. Section 5.2 describes the inclusion of the bridge sample data in development of the longitudinal scale.

Figure 5-2. Normal distributions of ability for adjacent samples, and difficulty parameters of common items: Mathematics (first grade, second grade bridge, and third grade): School years 1999–2000 and 2001–02



SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2000, second grade bridge study and spring 2002.

5.2 Development of the K-1-3 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 graders and the second grade bridge sample. The link between the K-1 and third grade assessment forms relied on 22 reading items and 14 mathematics items that were present in both versions of the assessments. These common items permitted the development of a vertical scale suitable for measuring gains in the early elementary years.

5.2.1 Evaluating Common Items

The first step in developing the longitudinal scale was evaluating the functioning of the common items at different time points. Although the content and presentation of each of the common

items were identical in the two versions of the assessments, 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 percent correct 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 first 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 third grade, when children could have had extensive practice in the skills tapped by "X," it could become much *easier* than the *same* set of "Y" computations. Such an item, showing a large difference in *relative* difficulty over time, should not be treated as a common item for the purpose of estimating gains.

In order to assess the common *functioning* of the overlapping reading and mathematics items, preliminary estimates of an IRT item and ability parameters were obtained, using all items in the K-1 and third 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 5-4 and 5-5 compare the actual with the predicted proportion correct for each of the 22 reading and 14 mathematics common items, respectively, based on the children who answered each of the items in each round of data collection, including the bridge sample. Table 5-6 shows the mean absolute discrepancy for each item averaged across the 6 samples, and for each sample averaged across all of the common items.

Note that the comparisons of actual vs. predicted percent correct for each question can be carried out *only for children who answered the question*. For questions that appeared in only one or two second-stage forms, or after a discontinue point in the routing test, these comparisons represent only a subset (and not a random subset) of the sample. For example, in table 5-4, reading items 137–140 were the last four items administered in the 20-item K-1 reading routing test. Although all children began the reading assessment with the same routing test, not all items were administered to all children. The routing test was discontinued at one of two points for children who had performed poorly on the first groups of items. The actual and predicted percent correct for items 137–140 shown in the table is the same or lower

for spring-kindergarten than for fall-kindergarten. This is a result of these items being administered to less than 4 percent of the fall-kindergarten group, but more than 16 percent of the spring-kindergarten children. Within each of these rounds, 47 percent of the children who *responded* to item 137 answered correctly. However, because of the discontinue rules, the correct respondents made up only 1.8 percent of the whole fall-kindergarten sample, but 8.3 percent of the whole spring-kindergarten group. A similar pattern appears for items 141–147, and for the same reason: these items were present only in the high second-stage form, which was taken by different percentages of the sample in each round. Thus the actual and predicted percentages in tables 5-4 and 5-5 should not be interpreted as population estimates, but as indicators of model fit based on the non-random subsets of children in each round who answered the questions.

For almost all of the items, the difference between the actual and predicted percent correct was very small, indicating common functioning of the items across time periods and good fit to the IRT model. Only one item, mathematics item 112, had an actual-predicted discrepancy that exceeded .10. Third graders were less successful in answering this item (69 percent correct) than had been predicted (85 percent correct). Half of the 22 reading common items and 8 of the remaining 13 mathematics common items had absolute discrepancies between actual and predicted proportion correct between .05 and .10 for one or more rounds of data collection. These discrepancies were too small to have serious impact on the ability estimates. All 22 reading common items and 13 of the 14 mathematics common items were treated as common items for calibrating final IRT parameters. The mathematics item with the largest discrepancy 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.

Table 5-4. Comparison of actual with predicted proportion correct, reading assessment common items, six data collections rounds: School years 1998–99, 1999–2000, and 2001–02

Reading assessment common item		Proportion correct																	
		Fall-kindergarten			Spring-kindergarten			Fall-first grade			Spring-first grade			Second grade bridge			Spring-third grade		
		Actual	Predicted	Diff.	Actual	Predicted	Diff.	Actual	Predicted	Diff.	Actual	Predicted	Diff.	Actual	Predicted	Diff.	Actual	Predicted	Diff.
Item 133	runs	.10	.10	.00	.29	.29	.00	.44	.44	-.01	.87	.84	.03	.98	.99	-.01	.99	.99	-.01
Item 134	down	.06	.06	.00	.18	.20	-.02	.33	.33	.00	.82	.78	.03	.97	.97	.00	.94	.96	-.02
Item 135	went	.08	.08	.00	.24	.23	.00	.36	.37	-.01	.81	.79	.02	.98	.98	-.01	.98	.99	-.01
Item 136	jeep	.08	.07	.01	.21	.21	.00	.32	.34	-.01	.77	.75	.02	.90	.95	-.05	.87	.93	-.07
Item 137	backpack	.47	.45	.03	.47	.45	.02	.54	.51	.02	.68	.68	-.01	.90	.92	-.02	.95	.94	.00
Item 138	ridebike	.47	.40	.07	.42	.39	.03	.45	.45	.00	.60	.61	-.01	.93	.90	.03	.92	.93	.00
Item 139	listen	.42	.38	.04	.39	.38	.01	.46	.45	.01	.63	.63	.00	.92	.92	.00	.94	.94	-.01
Item 140	sizes	.36	.33	.03	.35	.32	.03	.40	.39	.01	.56	.56	-.01	.90	.89	.00	.93	.93	.00
Item 141	through	.25	.19	.06	.17	.16	.01	.23	.21	.01	.38	.37	.01	.52	.54	-.02	.51	.54	-.04
Item 142	rage	.19	.14	.05	.12	.11	.02	.15	.15	-.01	.27	.28	-.01	.40	.41	-.01	.45	.42	.03
Item 143	toil	.21	.15	.06	.14	.12	.02	.17	.17	.00	.26	.28	-.01	.36	.39	-.03	.43	.39	.04
Item 144	capture	.15	.11	.04	.10	.09	.01	.11	.13	-.02	.22	.22	.00	.32	.31	.00	.34	.33	.01
Item 145	corner	.17	.14	.03	.12	.11	.00	.15	.15	-.01	.25	.26	-.01	.44	.37	.07	.41	.37	.04
Item 146	web	.18	.14	.04	.12	.12	-.01	.15	.16	-.01	.26	.26	.00	.40	.34	.06	.36	.34	.02
Item 147	strands	.08	.06	.03	.06	.05	.01	.06	.07	-.01	.11	.11	.00	.11	.14	-.03	.14	.14	-.01
Item 148	quiet	†	†	†	†	†	†	.26	.24	.01	.41	.41	.01	.87	.89	-.03	.91	.92	-.01
Item 149	weightless	†	†	†	†	†	†	.13	.13	.00	.22	.23	-.01	.79	.83	-.04	.90	.89	.01
Item 150	require	†	†	†	†	†	†	.14	.11	.02	.19	.19	.00	.68	.73	-.05	.80	.80	.00
Item 151	moisture	†	†	†	†	†	†	.32	.32	.00	.40	.43	-.03	.57	.62	-.04	.71	.69	.01
Item 152	preference	†	†	†	†	†	†	.28	.18	.10	.33	.26	.07	.22	.24	-.02	.32	.33	-.01
Item 153	ambition	†	†	†	†	†	†	.11	.08	.03	.22	.16	.06	.09	.12	-.04	.21	.21	.00
Item 154	criticism	†	†	†	†	†	†	.11	.10	.02	.15	.21	-.07	.11	.15	-.04	.27	.26	.01
Mean of means		.22	.19	.03	.23	.22	.01	.26	.25	.01	.43	.42	.00	.61	.62	-.01	.65	.65	.00

† Not applicable.

NOTE: Items 148-154 were part of a first grade supplemental form and were not administered in kindergarten. Actual and predicted proportions for each item are based only on children who answered the item. Proportions are unweighted.

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, and spring 2002.

Table 5-5. Comparison of actual with predicted proportion correct, mathematics assessment common items, six data collection rounds: School years 1998–99, 1999–2000, and 2001–02

Mathematics assessment common item		Proportion correct																	
		Fall-kindergarten			Spring-kindergarten			Fall-first grade			Spring-first grade			Second grade bridge			Spring-third grade		
		Actual	Predicted	Diff.	Actual	Predicted	Diff.	Actual	Predicted	Diff.	Actual	Predicted	Diff	Actual	Predicted	Diff.	Actual	Predicted	Diff.
Item 110	5 10 15 -- 25	.04	.06	-.02	.19	.20	-.01	.30	.34	-.04	.71	.68	.03	.95	.86	.09	.90	.85	.06
Item 111	5-1 oranges	.27	.24	.03	.43	.43	.00	.56	.56	.01	.79	.80	-.01	.86	.91	-.05	.82	.90	-.09
Item 112	5+2 marbles	.19	.17	.01	.38	.37	.01	.53	.50	.03	.75	.74	.00	.80	.86	-.06	.69	.85	-.16
Item 113	3+7 pennies	.10	.08	.02	.24	.26	-.01	.38	.40	-.02	.72	.72	.00	.87	.88	-.01	.85	.86	-.01
Item 114	1 3 -- 7 9	.09	.10	-.02	.14	.17	-.03	.21	.24	-.04	.50	.47	.03	.74	.72	.02	.81	.81	.00
Item 115	$2x5 lunch	.11	.06	.04	.14	.12	.02	.22	.19	.04	.39	.43	-.04	.68	.72	-.03	.82	.82	.00
Item 116	8/2 candies	.05	.03	.02	.08	.06	.02	.15	.11	.04	.26	.29	-.03	.53	.57	-.04	.72	.72	.00
Item 117	15/5=3 cars	.08	.05	.03	.11	.09	.02	.17	.15	.02	.35	.38	-.02	.62	.67	-.05	.78	.79	-.01
Item 118	to 12 by 2s	.31	.30	.01	.43	.44	-.01	.52	.55	-.04	.79	.79	.00	.91	.90	.01	.88	.88	-.01
Item 119	# heads	.23	.18	.05	.22	.21	.00	.26	.25	.00	.35	.38	-.03	.46	.42	.04	.47	.41	.06
Item 120	how many $	.19	.15	.05	.20	.18	.02	.25	.23	.02	.34	.38	-.04	.47	.43	.04	.50	.43	.07
Item 121	12-4 pennies	.12	.06	.06	.13	.09	.04	.16	.14	.02	.27	.30	-.03	.45	.36	.08	.38	.36	.02
Item 122	18-2-3 soccer	.05	.05	.00	.07	.07	.00	.10	.11	.00	.23	.24	-.01	.53	.45	.08	.62	.62	-.01
Item 123	4+4-2	.16	.17	-.01	.23	.22	.01	.30	.29	.02	.51	.50	.00	.56	.59	-.03	.53	.58	-.05
Mean of means		.14	.12	.02	.21	.21	.00	.29	.29	.00	.50	.51	.01	.67	.67	.01	.70	.71	.01

NOTE: Item 112 was removed from the common item list for final IRT calibration. Actual and predicted proportions for each item are based only on children who answered the item. Proportions are unweighted.

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, and spring 2002.

Table 5-6. Mean absolute discrepancies between actual and predicted performance, averaged over rounds and items, reading and mathematics assessments, six data collection rounds: School years 1998–99, 1999–2000, and 2001–02

Mean discrepancy by item, averaged over six rounds			
Reading		Mathematics	
Item 133	0.01	Item 110	0.04
Item 134	0.01	Item 111	0.03
Item 135	0.01	Item 112	0.05
Item 136	0.03	Item 113	0.01
Item 137	0.02	Item 114	0.02
Item 138	0.02	Item 115	0.03
Item 139	0.01	Item 116	0.02
Item 140	0.01	Item 117	0.03
Item 141	0.02	Item 118	0.01
Item 142	0.02	Item 119	0.03
Item 143	0.03	Item 120	0.04
Item 144	0.01	Item 121	0.04
Item 145	0.03	Item 122	0.02
Item 146	0.02	Item 123	0.02
Item 147	0.01		
Item 148	0.01		
Item 149	0.01		
Item 150	0.01		
Item 151	0.01		
Item 152	0.03		
Item 153	0.02		
Item 154	0.02		
Mean discrepancy by round, averaged over all common items			
Reading		Mathematics	
Round 1	0.02		0.03
Round 2	0.01		0.01
Round 3	0.01		0.02
Round 4	0.02		0.02
Bridge	0.03		0.05
Round 5	0.02		0.04
Total mean discrepancy	0.02		0.03

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), fall 1998, spring 1999, fall 1999, spring 2000, and spring 2002..

5.2.2 IRT Calibration and Scoring

IRT calibration was carried out using the PARSCALE program as described in chapter 3. Four decoding items in the reading assessment were deleted from the item pool after it was determined that their difficulty was so far above the level of the third grade sample that it was not possible to estimate stable item parameters for them. The reading estimation was based on the remaining 154 unique items that appeared in all forms of the reading assessments, including 22 reading items that were common to both the K-1 and third grade versions. Two mathematics items were deleted because of differential functioning for subgroups (see section 4.4.5), leaving 123 unique mathematics items in all assessment forms, including 13 common items linking the K-1 and third grade assessments. For each item, the IRT calibration resulted in a set of three item parameters that define a logistic function associated with the item. The height of the function at any point along an ability range corresponds to the estimated probability of a correct answer on the item for a person at that ability level.

The tables in appendix B present the results of the IRT calibration. The IRT a, b, and c parameters (for discrimination, difficulty, and guessing) are shown for each item, along with a "map" specifying the test form or forms in which the item appeared in each of the two versions of the assessment, the K-1 forms used in rounds 1-4, and the third grade forms for round 5. Fit statistics are included for the five rounds of data collection, showing the actual proportion correct (P+) for children who answered each item, the proportion predicted by the IRT model, and the difference between them. Because only a subset of items (one routing test and one second stage form) was administered to each child in each round, the number of cases on which the statistics are based is also included in the tables. The largest discrepancies between actual and predicted proportion correct tend to occur for rounds in which item statistics are based on relatively small amounts of data. For example, very few fall-kindergarten (round 1) children were routed to the high second stage reading or mathematics forms, and very few spring-first graders (round 4) received the lowest second stage forms. A few items in these sections for these rounds had absolute discrepancies of about .05 to .08 between actual and predicted proportion correct. Conversely, almost all of the items in the modal form for each round had absolute discrepancies of .03 or less. For the large majority of responses in each round, the fit of the IRT model to the data was very close.

Each of the rounds of data collection, kindergarten through third 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 3.1.1 for additional details. Table 5-7 presents theta (ability) means and standard deviations for the subpopulations of the reading and mathematics calibrations. The theta estimates are standardized to mean = 0.0 and standard deviation = 1.0 for all rounds combined.

Table 5-7. IRT theta (ability) means and standard deviations by subpopulation, six data collection rounds: School years 1998–99, 1999–2000, and 2001–02

Round	Reading		Mathematics	
	Mean	SD[1]	Mean	SD
All rounds combined	0.00	1.00	0.00	1.00
Round 1 (fall-kindergarten)	-1.06	0.59	-1.01	0.61
Round 2 (spring-kindergarten)	-0.43	0.57	-0.42	0.58
Round 3 (fall-first grade)	-0.17	0.58	-0.12	0.59
Round 4 (spring-first grade)	0.50	0.52	0.48	0.53
Second grade bridge sample	1.13	0.33	1.03	0.38
Round 5 (spring-third grade)	1.30	0.39	1.30	0.48

[1] Standard deviation.
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), fall 1998, spring 1999, fall 1999, spring 2000, and spring 2002.

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 representing an estimate of the number of items the student would have answered correctly if he or she had taken all 154 reading items, or all 123 mathematics items. (The same procedure applies to the 62 items in the science assessment, which was given in third grade only.) 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. Eight such levels were defined for reading, and seven for

mathematics, based on items from the K-1 and third 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. 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 correct from item-response-based cluster scores (1 or 0) were observed for all levels in all rounds, for both reading and mathematics.

The parameters for the eight reading and seven mathematics proficiency levels are shown in table 5-8. The very high "a" parameters are consistent with the assumption that 4-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 student to pass a 4-item cluster by guessing; the guessing parameters (c) were all fixed at zero.

Table 5-8. IRT parameters for reading and mathematics proficiency levels, based on items from kindergarten, first grade, and third grade assessments: School years 1998–99, 1999–2000, and 2001–02

Proficiency	Reading			Mathematics		
	a	b	c	a	b	c
Level 1	4.25	-1.43	0.0	3.09	-1.96	0.0
Level 2	2.89	-0.77	0.0	2.68	-1.11	0.0
Level 3	2.75	-0.45	0.0	3.79	-0.49	0.0
Level 4	4.25	0.13	0.0	3.19	0.20	0.0
Level 5	5.80	0.59	0.0	3.95	0.90	0.0
Level 6	6.77	1.01	0.0	5.22	1.43	0.0
Level 7	5.31	1.36	0.0	5.88	1.81	0.0
Level 8	3.28	1.58	0.0	†	†	†

† Not applicable.
NOTE: Only seven mathematics proficiency levels were defined.
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, and spring 2002.

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 A32 and A33 in appendix A present subgroup differences with respect to mastery of the level that represents the modal "highest level" score within each round.

5.3 Applications

This section describes issues in selection and use of scores for analyzing status and gain in cognitive skills. Appendix A includes breakdowns by gender, ethnicity, SES, and school type for all of the third grade direct cognitive measures. For measures that can be compared with the analogous scores in earlier rounds, results for rounds 1 through 4 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.

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 demographics, 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 may be used 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, they 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 third grade may be working on learning to make evaluative judgments based on reading material, which would show up as large gains in reading level 8. Less skilled readers may show their largest gains between first and

third grade at levels 5 or 6, comprehension of words in context and literal inference. 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 proficient level mastered also, in theory, is derived from item responses, although a relatively small number of IRT-based estimates were substituted for missing data.

Routing test number-right scores for the third grade reading, math, and science assessments are based on 15, 17, and 15 items respectively (20, 16, and 12 items for the K-1 grade 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 third 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 grade routing tests were based on different and more difficult sets of items. The third grade routing test number-right scores should *not* be compared with the kindergarten or first grade routing test number-right scores.

Item cluster scores in reading (e.g., Decoding Score Gr 3) and science (e.g., Life Science Gr 3) 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 five items in each of these item cluster scores, reliabilities tend to be relatively low.

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. For about 20 percent of these scores, 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.

5.3.2 Notes on Measuring Gains

This section outlines approaches to measuring gains that rely on multiple criterion-referenced points to identify different patterns of student 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 total scale score points in reading between fall- and spring-kindergarten shows an average increase of about 10 points and gains of about 40 points between spring-first grade and spring-third 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 30 points between first and third grade, again starting from very different initial status.

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, who are achieving their 10-point gains by learning to read words and sentences.

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 students 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 eight reading and seven 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 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 A29). 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 14 percent at the end of third grade (shown in table A31). 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 time 1 and time 2 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 student 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).

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 September 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 impact 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 *ECLS-K, Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05).

This page is intentionally left blank.

6. PSYCHOMETRIC CHARACTERISTICS OF THE INDIRECT MEASURES AND THE DIRECT SELF DESCRIPTION QUESTIONNAIRE

Chapter 2 describes the selection and development of the third grade indirect measures and the Self-Description Questionnaire. The indirect measures were teacher evaluations of children's academic and social skills. The Self-Description Questionnaire asked children to rate their competence and interest in school subjects and their behavior and relationships with peers. This chapter provides details of the psychometric characteristics of these instruments. In addition, the relationships between the direct and indirect cognitive measures are explored.

6.1 Teacher Measures

In the spring-third grade data collection (round 5), teachers of the sampled children were asked to evaluate each child's academic and social skills. The third grade teacher measures were similar in design to those administered in kindergarten and first grade, and shared some common items with the earlier instruments. Teachers were instructed to rate children's current skills and behaviors according to grade-level expectations. The resulting third grade scores, while sharing names with the kindergarten and first grade measures collected earlier, are scaled differently. They should not be directly compared with kindergarten and first grade scores for the purpose of evaluating gains over time. Data collected in the earlier rounds may, however, be used as covariates in analyzing third grade achievement and behavioral data. Details of the kindergarten and first grade teacher measures (and similar behavioral ratings provided by parents) may be found in *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05).

Differential item functioning (DIF) analysis was not carried out for the indirect measures. DIF analysis of the Academic Rating Scale (ARS) was not appropriate for several reasons. First, the ratings were produced by the teacher, not by direct observation of the child. Therefore, there is a confounding source of difference, namely the teacher's attitudes or potential bias, that cannot be separated from the child's performance. Second, even if it could be determined that teachers' ratings were completely accurate and unbiased, DIF would also be impossible for the ARS because there is no satisfactory criterion for matching. DIF analysis depends on the assumption that, for subsets of individuals *matched on overall ability level*, performance on each test item should be about the same. The

ARS scales are too short to provide a matching criterion (i.e., each item represents too big a part of the total score needed for matching), and there is no independent measure of the same construct that could be used for this purpose. The direct cognitive score would not be an appropriate criterion because the ARS includes process questions that are not represented in the direct cognitive tests. Third, 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 is not a psychometric characteristic of the scale itself.

DIF analysis of the Social Rating Scale (SRS) was not carried out because DIF assumptions are not relevant to behavioral and attitudinal measures. The basic premise of DIF is that for subsets of individuals matched according to a criterion (such as a score on the total set of items or some external criterion), *similar item performance* for different subgroups should be observed. Significant deviation from this could indicate that an item is measuring differently for different groups. For behavioral measures such as SRS, there can be no expectation that ratings *should* be the same for different groups. Any group differences in ratings may reflect either legitimate real differences in the group's attitude or behavior on an item or set of items, or factors having to do with the standards or attitudes of the rater (teacher), not differential functioning or flaws in the items.

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 social and academic ratings assigned to children. Secondary analysis of these relationships may reveal differences in the standards used in the academic (ARS) and social (SRS) ratings.

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

The ARS evaluated achievement in the three domains that are also assessed in the direct cognitive assessment battery: language and literacy (reading), mathematical thinking, and science. In addition, the ARS provided the only assessment of the child's skills, knowledge, and behavior in social studies. For each of the four scales, the teacher was asked if he or she was the child's primary teacher in the area. If not, the teacher was instructed to consult with the person most familiar with the child's progress in the subject area before assigning the ratings.

The ARS was designed both to overlap and to 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 importantly, the ARS includes items designed to measure both the process and products of children's learning in school, whereas the direct cognitive battery assesses only the products of children's achievement. The scope of curricular content represented in the indirect measures was designed to be broader than the content represented on the direct cognitive measures. Unlike the direct cognitive measures, which were designed to measure gain on a longitudinal scale spanning kindergarten entry through the end of fifth grade, the ARS is targeted to a specific grade level. The questions range from criterion-referenced items (e.g., "shows understanding of place value with whole numbers") to others with a more norm-referenced point of view (e.g., "reads third grade books [fiction] independently with comprehension"). Teachers evaluating the children's skills were instructed to rate each child compared with other children of the same age/grade level. Response options for each item ranged from 1 ("not yet") to 5 ("proficient"). See section 2.3 for additional details on the design and development of the ARS instrument.

The one-parameter IRT model (Rasch model) used to estimate ARS scores is described in detail in chapter 3. The reliability for each of the scales is very high (table 6-1). The summary fit statistics for persons and items are acceptable for all the scales (table 6-2). The fit statistics for the step calibrations indicate that the lowest category ("Not yet") was used less than expected.

Table 6-1. Academic Rating Scale (ARS) person reliability for the Rasch-based score, spring-third grade: School year 2001–02

Scale	Reliability
Language and Literacy	.95
Mathematical Thinking	.94
Science	.95
Social Studies	.93

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

Table 6-2. Academic Rating Scale (ARS) fit statistics for persons and items, spring-third grade: School year 2001–02

Scale	Infit MNSQ[1]	Outfit MNSQ
Persons		
Language and Literacy	1.00	1.00
Mathematical Thinking	1.02	1.02
Science	.96	.96
Social Studies	.98	.98
Items		
Language and Literacy	1.00	1.00
Mathematical Thinking	1.06	1.10
Science	1.00	.97
Social Studies	1.01	1.00

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

The ARS scores were scaled to have a low of "1" and a high of "5" to correspond to the 5-point rating scale that teachers used in rating children on these items, but they should not be interpreted as mean scores. The item difficulties and student scores are placed on a common scale. Students have a high probability of receiving a high rating on items below their scale score and a lower probability of receiving a high rating on items above their scale score. For example, a child whose Rasch IRT scale score is 4.0 would have a greater than 50 percent probability of having received a rating of "5" on all items whose difficulty is below 4.0 on the scale. Students who received maximum ratings on all the items or minimum ratings on all the items were assigned an estimated score.

The ARS scales were designed to provide information on children's abilities at a given point in time, rather than provide a measure of change over time. The sets of items developed for the third grade ARS ratings were different from the items used in the kindergarten and first grade instruments. Although the third grade item stems have some similarities to those used in the earlier forms, the extended item descriptions include grade-appropriate performance criteria that describe the level of proficiency a child should have reached in order to receive the highest rating. For example, "uses a variety of strategies to solve math problems" appeared in all versions of the ARS, kindergarten through third grade. In fall-kindergarten this item was described as "using manipulative materials, looking for a pattern, or acting out a problem" while the third grade ARS described the same stem as "adds 100 and then subtracts 4 when doing the mental math problem 467+96, or writes the algorithms or equations needed to solve a word problem, or orders steps sequentially in a multistep problem." Obviously, a fall-kindergarten rating with respect to the first description does not represent the same level of skill as the same rating based on the

third grade criterion. As a result, the ARS score metric is different at each point in time, and change scores should *not* be used to compare third grade ratings with those from earlier rounds. Covariance models may be used to compare teachers' ratings of performance in different grades. Before using these variables in such analyses, the distribution of the samples should be assessed to determine if the assumption of normal distribution is met.

On the ARS, teachers indicated "not applicable" when the knowledge, skill, or behavior has not been introduced to the classroom. Because some children might already have had this skill (from home or other opportunities for learning), the "not applicable" ratings were treated as missing data and the child's score was estimated based on the items on which the child was rated. Although the Rasch program estimates scores for all children based on the information provided, scores estimated on a limited number of responses are less reliable than scores with more ratings. ARS scores were computed only if at least 60 percent of the items in the scale were given ratings. In other words, if more than 40 percent of the items in a scale were not rated, then the score was set to missing.

The weighted means and standard deviations for the third grade ARS scores are shown in table 6-3. Score breakdowns for population subgroups are presented in tables 6-18 through 6-21 at the end of this chapter.

Table 6-3. Academic Rating Scale (ARS) means and standard deviations, spring-third grade: School year 2001–02

Scale	Weighted mean	Standard deviation
Language and Literacy	3.27	0.89
Mathematical Thinking	3.08	0.75
Science	3.17	0.93
Social Studies	3.02	0.85

NOTE: Table estimates are based on C5CW0 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 2002.

6.1.1.1 Floor and Ceiling

As noted in the section on the development of the ARS, the criteria for some of the items was 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 6-4 presents the percentage of children at the ceiling and floor of the measures. The percentages of perfect scores are somewhat lower than had been found for the same scales in the first grade ARS, although the percentages of minimum scores were comparable to earlier rounds.

Table 6-4. Percent of sample with perfect and minimum Academic Rating Scale (ARS) scores, spring-third grade: School year 2001–02

Description	Percent
Perfect scores	
Language and Literacy	6.0
Mathematical Thinking	4.7
Science	6.7
Social Studies	5.2
Minimum scores	
Language and Literacy	0.8
Mathematical Thinking	0.7
Science	1.4
Social Studies	0.2

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 2002.

Tables 6-5 to 6-8 provide the estimates of difficulty for each of the items. Higher difficulty values mean that teachers rated fewer students as proficient on those items. The items on each of the measures tended to cluster in difficulty.

Table 6-5. Academic Rating Scale (ARS) language and literacy item difficulties (arranged in order of difficulty), spring-third grade: School year 2001–02

Item difficulty	Item number and abbreviated content
2.69	Q3. Conveys ideas clearly when speaking
2.72	Q6. Reads third grade books (fiction) independently with comprehension
2.74	Q5. Reads fluently
2.87	Q4. Uses various strategies to gain information
3.03	Q7. Reads and comprehends expository text
3.19	Q10. Makes mechanical corrections when reviewing a rough draft
3.22	Q8. Composes multi-paragraph stories/reports
3.28	Q9. Rereads and reflects on writing, making changes to clarify or elaborate

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

Table 6-6. Academic Rating Scale (ARS) mathematical thinking item difficulties (arranged in order of difficulty), spring-third grade: School year 2001–02

Item difficulty	Item number and abbreviated content
2.45	Q7. Shows understanding of place value with whole numbers
2.53	Q3. Creates and extends patterns
2.71	Q5. Recognizes properties of shapes and relationships among shapes
2.73	Q9. Surveys, collects, and organizes data into simple graphs
2.78	Q4. Uses a variety of strategies to solve math problems
2.81	Q6. Uses measuring tools accurately
2.83	Q8. Makes reasonable estimates of quantities and checks answers
3.16	Q10. Models, reads, writes, and compares fractions
3.69	Q11. Divides a 3-digit number by a 1-digit number

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

Table 6-7. Academic Rating Scale (ARS) science item difficulties (arranged in order of difficulty), spring-third grade: School year 2001–02

Item difficulty	Item number and abbreviated content
2.77	Q5. Classifies and compares living and non-living things in different ways
2.84	Q3. Makes logical predictions when conducting scientific investigations
2.87	Q8. Demonstrates understanding of life science concepts
2.93	Q9. Demonstrates understanding of earth and space science concepts
2.97	Q6. Forms explanations and conclusions based on observation and investigation
3.01	Q7. Demonstrates understanding of physical science concepts
3.07	Q4. Communicates scientific information

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

Table 6-8. Academic Rating Scale (ARS) social studies item difficulties (arranged in order of difficulty), spring-third grade: School year 2001–02

Item difficulty	Item number and abbreviated content
2.56	Q7. Knows how to use maps and globes to locate and derive information
2.67	Q3. Identifies similarities and differences in group habits and living patterns
2.84	Q6. Recognizes the reciprocal influence of environment on people
2.88	Q5. Demonstrates understanding of the ways in which the past influences the present
3.19	Q4. Shows understanding of the purpose and structure of government functions
3.33	Q8. Demonstrates understanding of the U. S. economic system

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

Tables 6-9 to 6-12 provide standard errors (SE) for each of the ARS Rasch scores for third grade. The "Score" column is the sum of the raw score ratings. "Measure" is the Rasch-based score. 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 6-9. Academic Rating Scale (ARS) language and literacy standard errors, spring-third grade: School year 2001–02

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
8	1.00E	.45	19	2.45	.15	30	3.56	.15
9	1.32	.26	20	2.54	.15	31	3.66	.15
10	1.53	.20	21	2.64	.15	32	3.75	.15
11	1.68	.18	22	2.74	.16	33	3.85	.15
12	1.80	.16	23	2.84	.16	34	3.95	.16
13	1.90	.16	24	2.94	.16	35	4.06	.16
14	2.00	.15	25	3.05	.16	36	4.17	.17
15	2.09	.15	26	3.15	.16	37	4.30	.18
16	2.18	.15	27	3.26	.16	38	4.46	.21
17	2.27	.15	28	3.36	.16	39	4.68	.27
18	2.36	.15	29	3.46	.16	40	5.00E	.45

NOTE: E = estimated extreme score. The "Score" column is the sum of the raw score ratings. "Measure" is the Rasch-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 2002.

Table 6-10. Academic Rating Scale (ARS) mathematical thinking standard errors, spring-third grade: School year 2001–02

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
9	1.00E	.49	22	2.46	.14	35	3.41	.14
10	1.34	.28	23	2.53	.14	36	3.49	.15
11	1.55	.21	24	2.60	.14	37	3.57	.15
12	1.69	.18	25	2.67	.14	38	3.66	.15
13	1.79	.16	26	2.74	.14	39	3.75	.16
14	1.88	.15	27	2.82	.14	40	3.85	.17
15	1.96	.14	28	2.89	.14	41	3.97	.18
16	2.04	.14	29	2.96	.14	42	4.11	.20
17	2.11	.14	30	3.03	.14	43	4.29	.24
18	2.18	.13	31	3.11	.14	44	4.58	.32
19	2.25	.13	32	3.18	.14	45	5.00E	.52
20	2.32	.13	33	3.26	.14	†	†	†
21	2.39	.13	34	3.33	.14	†	†	†

† Not applicable.
NOTE: E = estimated extreme score. The "Score" column is the sum of the raw score ratings. "Measure" is the Rasch-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 2002.

Table 6-11. Academic Rating Scale science (ARS) standard errors: School year 2001–02

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
7	1.00E	.40	17	2.36	.15	27	3.63	.19
8	1.29	.23	18	2.47	.15	28	3.80	.20
9	1.48	.16	19	2.58	.16	29	3.97	.19
10	1.61	.15	20	2.71	.17	30	4.12	.17
11	1.73	.15	21	2.85	.18	31	4.25	.16
12	1.83	.15	22	2.99	.17	32	4.37	.17
13	1.94	.15	23	3.12	.16	33	4.51	.18
14	2.05	.15	24	3.24	.16	34	4.71	.24
15	2.15	.15	25	3.36	.16	35	4.99E	.40
16	2.26	.15	26	3.48	.17	†	†	†

† Not applicable.
NOTE: E = estimated extreme score. The "Score" column is the sum of the raw score ratings. "Measure" is the Rasch-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 2002..

Table 6-12. Academic Rating Scale (ARS) social studies standard errors, spring-third grade: School year 2001–02

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
6	1.00E	.45	15	2.47	.17	24	3.71	.19
7	1.33	.27	16	2.59	.17	25	3.86	.19
8	1.56	.21	17	2.72	.17	26	4.02	.20
9	1.73	.19	18	2.84	.17	27	4.19	.21
10	1.87	.18	19	2.97	.18	28	4.39	.23
11	2.00	.17	20	3.11	.18	29	4.65	.28
12	2.12	.17	21	3.25	.19	30	5.00E	.46
13	2.23	.17	22	3.40	.19	†	†	†
14	2.35	.17	23	3.55	.19	†	†	†

† Not applicable.
NOTE: E=estimated extreme score. The "Score" column is the sum of the raw score ratings. "Measure" is the Rasch-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 2002.

The majority of teachers rated more than one student on the ARS. The number of students rated by each teacher ranged from one to more than 20. 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 as many ECLS-K sample children as they had in class.

6.1.2 Social Rating Scale (SRS)

The Social Rating Scale (SRS) is an adaptation of the Social Skills Rating System (Gresham and Elliott, 1990). As part of a self-administered questionnaire, third grade teachers were asked to judge how often students exhibited certain social skills and behaviors. (In kindergarten and first grade, SRS questions had been asked of both teachers and parents.) Teachers used a frequency scale to report on how often the student demonstrated the social skill or behavior described (1 = never to 4 = very often). The 24 SRS items used in kindergarten and first grade were included in the third grade SRS, and two new items were added.

Five teacher SRS scales, with the same names as the kindergarten and first grade SRS scales, were computed based on responses to the items. The scales are the following: Approaches to Learning, Self-Control, Interpersonal Skills, Externalizing Problem Behaviors, and Internalizing Problem Behaviors. Two items were added to the third grade scales due to a high number of maximum scores on the third grade field test. One item was added to the Externalizing Problem Behavior scale ("child talks during quiet study time"). The other additional item "child follows classroom rules" was added to the SRS in an attempt to increase variance in the self-control scale. Analysis of the item responses indicated that it contributed strongly to the Approaches to Learning scale, increasing the variance and reliability of that scale. Thus, this item was included in the Approaches to Learning scale.

In third grade, examination of the responses suggested a different perception of a student's self-control and interpersonal social abilities. The Self-Control scale includes items on control of attention as well as control of emotions and behavior in interactions. Third grade students who were rated higher on self-control were also rated higher on interpersonal skills that involved peers. Thus, in addition to the Self-Control and Interpersonal social abilities scale scores, a Peer Relations scale score was included. This additional scale combines responses on both the interpersonal and self-control scale items that relate to peers.

Although 24 of the 26 third grade SRS items were the same as items in the kindergarten-first grade (K-1) instrument, teachers may place different interpretations on the meaning of the items at different time points. Therefore these scores would be most appropriately used as covariates rather than as change scores.

The score on each SRS scale is the mean of ratings on the items included in the scale. Scores were computed only if the student was rated on at least two-thirds of the items in that scale. Factor analyses were used to confirm the scales. The split-half reliabilities for the teacher SRS scales were high (table 6-13).

Table 6-13. Split-half reliability for the teacher Social Rating Scale (SRS) scores, spring-third grade: School year 2001–02

Scale	Split-half reliability
Approaches to Learning	.91
Self-Control	.79
Interpersonal	.89
Externalizing Problem Behaviors	.89
Internalizing Problem Behaviors	.76
Peer Relations (Self-Control and Interpersonal Combined)	.92

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

Weighted means and standard deviations for these scales are shown in table 6-14. About 90 percent of the children whose teachers provided social ratings data were in third grade during the round 5 data collection, and about 9 percent were in first or second grade. Numbers in the table are for third graders, with scores for children who at round 5 were still in first or second grade shown in parentheses. The number of children who had advanced to fourth or fifth grade by round 5 was too small to be analyzed separately. SRS score statistics for subpopulations are presented in tables 6-22 through 6-27 at the end of this chapter, with scores for third graders shown separately from those of children in first and second grade.

Table 6-14. Teacher Social Rating Scale (SRS) score means and standard deviations, spring-third grade: School year 2001–02

Description	Weighted mean	Standard deviation
Approaches to Learning	3.1 (2.7)	0.7 (0.7)
Self-Control	3.2 (3.0)	0.6 (0.7)
Interpersonal	3.1 (2.8)	0.6 (0.7)
Externalizing Problem Behaviors	1.7 (1.9)	0.6 (0.7)
Internalizing Problem Behaviors	1.6 (1.8)	0.5 (0.6)
Peer Relations (Self-Control and Interpersonal Combined)	3.2 (2.9)	0.6 (0.6)

NOTE: Table estimates based on C5CW0 weight. Numbers outside of parentheses represent children in third grade at the time of assessment. Numbers within parentheses represent first and second graders at the time of assessment. The range of possible 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 2002.

Care should be taken when entering these scales into the same analysis due to problems of multicollinearity. The intercorrelations among the five independent SRS factors (that is, excluding the combined peer relations scale) are generally high. Absolute values of correlations among the Approaches to Learning, Self-Control, Interpersonal Skills, and Externalizing Problem Behaviors scales range from .59 to .81 for third graders. Only the Internalizing Problem Behaviors scale had substantially weaker relationships with the other measures, with correlations of .32 to .41. Patterns of correlations for children who were still in first or second grade in the third grade round were very similar to patterns for the on-grade-level children, and were also consistent with results in the kindergarten and first grade rounds.

6.2 Self-Description Questionnaire (SDQ)

For the first time in the ECLS-K, children were asked to provide self-assessments of their academic and social skills. In the SDQ, third grade students rated their perceived competence and interest in reading, mathematics, and all school subjects.[1] They also rated their perceived competence and popularity with peers and reported on problem behaviors with which they might struggle. The Externalizing Problems scale included questions about anger and distractibility, while the Internalizing Problems scale included items on sadness, loneliness, and anxiety. For further detail on the development and content of the SDQ, see chapter 2. Students rated whether each item was "not at all true," "a little bit true," "mostly true," or "very true." Five scales were produced from the SDQ items. The scale scores on

[1] The SDQ was adapted, with permission, from the *Self-Description Questionnaire-I* (Marsh, 1990). See chapter 2.

all SDQ scales represent the mean rating of the items included in the scale. Students who responded to the SDQ answered virtually all of the questions, so treatment of missing data was not an issue. As with most measures of social-emotional behaviors, the distributions on these scales are skewed (negatively skewed for the positive social behavior scales, and positively skewed for the problem behavior scales). The reliability is lower for scales with only six items (see table 6-15). Weighted means and standard deviations for these scales are shown in table 6-16.

Table 6-15. Self-Description Questionnaire (SDQ) scale reliabilities, spring-third grade: School year 2001–02

Description	Number of items	Alpha coefficient
Perceived Interest/Competence — Reading	8	.87
Perceived Interest/Competence — Math	8	.90
Perceived Interest/Competence — All Subjects	6	.79
Perceived Interest/Competence — Peer Relations	6	.79
Externalizing Problems	6	.77
Internalizing Problems	8	.81

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

Table 6-16. Self-Description Questionnaire (SDQ) weighted means and standard deviations, spring-third grade: School year 2001–02

Description	Weighted mean	Standard deviation
Perceived Interest/Competence — Reading	3.26	0.66
Perceived Interest/Competence — Math	3.16	0.79
Perceived Interest/Competence — All Subjects	3.03	0.65
Perceived Interest/Competence — Peer Relations	2.92	0.66
Externalizing Problems	2.02	0.71
Internalizing Problems	2.22	0.74

NOTE: Table estimates based on C5CW0 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 2002.

SDQ score statistics for subpopulations are presented in tables 6-28 through 6-33 at the end of this chapter.

6.3 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 cause and effect.) More complete discussions of construct validity may be found in Campbell and Fiske (1959) and Campbell (1960).

Correlations among thirteen third grade measures were examined for evidence of convergent and discriminant validity. These measures included four teacher ratings of children's achievement (ARS), three selected teacher ratings of children's attitudes and behaviors (SRS), three children's self-ratings of achievement (SDQ), and direct cognitive scores in the three subject areas assessed. These correlations are shown in table 6-17. The thirteen measures are as follows:

1. ARS Lit — Teacher ARS score for Language and Literacy
2. ARS Math — Teacher ARS score for Mathematical Thinking
3. ASR Sci — Teacher ARS score for Science
4. ARS Soc — Teacher ARS score for Social Studies
5. AppLearn — Teacher SRS factor score for Approaches to Learning
6. InterPers — Teacher SRS factor score for Interpersonal
7. SelfCon — Teacher SRS factor score for Self-Control
8. SDQ Read — Child's self-rating of competence in reading
9. SDQ Math — Child's self-rating of competence in math
10. SDQ All — Child's self-rating of competence in all subjects
11. ReadThet — Direct cognitive test theta (ability) estimate for Reading

12. MathThet Direct cognitive test theta (ability) estimate for Mathematics

13. SciThet Direct cognitive test theta (ability) estimate for Science

Table 6-17. Intercorrelations among the indirect cognitive teacher ratings (ARS), selected teacher socio-behavioral measures (SRS), selected child self-ratings (SDQ), and direct cognitive test scores, spring-third grade: School year 2001–02

	Round 5												
Measures	ARS Lit	ARS Math	ARS Sci	ARS Soc	SRS App Learn	SRS Inter Pers	SRS Self Con	SDQ Read	SDQ Math	SDQ All	Read Thet	Math Thet	Sci Thet
ARS Lit.	1.00												
ARS Math	.82	1.00											
ARS Sci	.79	.83	1.00										
ARS Soc	.77	.80	.77	1.00									
SRSAppLearn	.60	.49	.47	.45	1.00								
SRSInterPers	.37	.31	.30	.28	.71	1.00							
SRSSelfCon	.30	.26	.24	.22	.68	.81	1.00						
SDQ Read	.21	.11	.12	.11	.16	.09	.07	1.00					
SDQ Math	.03	.11	.06	.04	.10	.04	.03	.18	1.00				
SDQ All	.11	.09	.08	.06	.20	.12	.10	.53	.55	1.00			
Read Thet	.65	.53	.53	.50	.42	.25	.22	.18	-.09	-.01	1.00		
Math Thet	.59	.59	.53	.49	.39	.21	.19	.04	.11	.01	.73	1.00	
Sci Thet	.50	.45	.49	.42	.29	.17	.17	.05	-.05	-.06	.72	.72	1.00

NOTE: Table estimates based on C5CW0 weight.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002.

Indirect ARS measures 1 to 3 have counterparts in measures 11 to 13, 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).

The correlation of the ARS language/literacy measure with ARS mathematical thinking is .82; the comparable correlation for the direct cognitive measure of reading with mathematics is .73. The correlation between the direct reading and mathematics scores continues a slight but steady decline that began in the kindergarten rounds, suggesting the possibility of some divergence of the two skills over time. During the same interval, the corresponding correlations for ARS remained consistently high. The correlations of the ARS science scale with the language/literacy and mathematical thinking measures are

also high (.79 and .83, respectively), while the direct cognitive correlations of reading and mathematics with science are somewhat lower (.72 for both). These patterns are consistent with correlations observed for the kindergarten and first grade measures. The differences between the two sets of correlations suggest somewhat less discriminant validity for the ARS than for the direct measures.

When one examines the cross-correlations from a convergent validity perspective, differences between the indirect and direct measures are also found. One would expect that the ARS score in each subject area would be more closely related to the direct measure of the same subject than to measures of the other subjects. This is true for language/literacy (with direct reading) and mathematical thinking (with direct math), 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 were almost exactly the same as those with the direct mathematics score. In third grade, the ARS science scale was 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. The same pattern had been observed in kindergarten and first grade, with general knowledge scores (instead of science) failing to show the highest correlations between indirect and direct measures of the corresponding subject area. The finding of relatively low convergent validity is a consequence of the high correlations among the indirect cognitive measures, .77 to .83. Correlations this high mean that the measures are unlikely to show strong differential relationships with other variables, even those designed to assess similar constructs.

The indirect cognitive measures also show consistently higher relationships with behavioral scales such as teacher ratings of approaches to learning, interpersonal behavior, and self-control than do the comparable direct cognitive measures (table 6-17). The higher intercorrelations for the indirect cognitive measures may be partly due to the fact that they do indeed measure process in addition to products. Teachers' views of children's attitudes and behavior may also influence their ratings of all content domains. However, regardless of the reason(s) for the greater "halo" effect, one is less likely to find differential relationships with other external process or skill measures for the indirect ratings compared with the direct measures. An additional consequence of having a significant part of the "halo" effect coming from the sharing of the learning process variable "approaches to learning" (especially for the language/literacy scale, $r = .60$) is that the indirect cognitive scale scores are somewhat more difficult to interpret. The same teacher rating score may represent differential contributions of achievement vs. behavioral factors for different children.

Correlations of children's self-ratings with other measures were low. Only the self-rating of reading competence with the teacher rating of language/literacy and the self-rating of competence in all school subjects with the teacher rating of approaches to learning, reached correlations of .20. This suggests that children use different criteria than teachers use when rating academic competence. Teachers are more knowledgeable about national standards and had more specific criteria to use when rating academic competence. Children's self-perceptions reflect not only the feedback that they receive from others about their performance, but may also be 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). As noted earlier, score breakdowns for population subgroups are presented in tables 6-18 through 6-21.

Table 6-18. Score breakdown, Academic Rating Scale (ARS), language and literacy, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total	11,268	3.27	0.89
Sex			
Male	5,644	3.16	0.89
Female	5,624	3.39	0.87
Race/ethnicity			
White, non-Hispanic	6,886	3.38	0.87
Black, non-Hispanic	1,339	2.97	0.89
Hispanic, race specified	867	3.21	0.90
Hispanic, race not specified	869	3.08	0.88
Asian	685	3.50	0.87
Hawaiian, other Pacific Islander	148	3.15	0.80
American Indian/Alaska Native	181	2.91	0.81
More than one race, non-Hispanic	281	3.31	0.82
Socieconomic status			
First quintile (lowest)	1,379	2.83	0.90
Second quintile	1,769	3.11	0.86
Third quintile	1,982	3.26	0.84
Fourth quintile	2,252	3.47	0.82
Fifth quintile (highest)	2,603	3.70	0.78
School type			
Public school	8,882	3.25	0.90
Private school	2,383	3.41	0.80

[1] Standard deviation.
NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-19. Score breakdown, Academic Rating Scale (ARS), mathematical thinking, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	11,095	3.08	0.75
Sex			
Male	5,591	3.09	0.76
Female	5,504	3.08	0.74
Race/ethnicity			
White, non-Hispanic	6,769	3.15	0.74
Black, non-Hispanic	1,312	2.89	0.74
Hispanic, race specified	863	3.01	0.74
Hispanic, race not specified	870	2.99	0.76
Asian	679	3.33	0.78
Hawaiian, other Pacific Islander	144	2.97	0.68
American Indian/Alaska Native	176	2.81	0.58
More than one race, non-Hispanic	269	3.09	0.72
Socioeconomic status			
First quintile (lowest)	1,362	2.76	0.75
Second quintile	1,739	2.96	0.75
Third quintile	1,959	3.07	0.70
Fourth quintile	2,205	3.22	0.70
Fifth quintile (highest)	2,558	3.40	0.70
School type			
Public school	8,762	3.07	0.76
Private school	2,330	3.16	0.67

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-20. Score breakdown, Academic Rating Scale (ARS), science, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	10,693	3.17	0.93
Sex			
Male	5,377	3.16	0.95
Female	5,316	3.17	0.91
Race/ethnicity			
White, non-Hispanic	6,599	3.28	0.92
Black, non-Hispanic	1,283	2.88	0.94
Hispanic, race specified	793	3.05	0.93
Hispanic, race not specified	824	2.94	0.91
Asian	607	3.36	0.91
Hawaiian, other Pacific Islander	148	2.97	0.86
American Indian/Alaska Native	173	2.83	0.82
More than one race, non-Hispanic	254	3.23	0.86
Socioeconomic status			
First quintile (lowest)	1,332	2.72	0.91
Second quintile	1,676	3.01	0.92
Third quintile	1,888	3.17	0.88
Fourth quintile	2,140	3.37	0.87
Fifth quintile (highest)	2,433	3.56	0.87
School type			
Public school	8,518	3.15	0.94
Private school	2,172	3.29	0.88

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-21. Score breakdown, Academic Rating Scale (ARS), social studies, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	10,661	3.02	0.85
Sex			
Male	5,343	2.99	0.87
Female	5,318	3.05	0.83
Race/ethnicity			
White, non-Hispanic	6,541	3.10	0.85
Black, non-Hispanic	1,291	2.87	0.87
Hispanic, race specified	810	2.91	0.84
Hispanic, race not specified	823	2.84	0.82
Asian	612	3.15	0.82
Hawaiian, other Pacific Islander	146	2.86	0.78
American Indian/Alaska Native	175	2.68	0.74
More than one race, non-Hispanic	251	3.04	0.81
Socioeconomic status			
First quintile (lowest)	1,311	2.67	0.84
Second quintile	1,675	2.87	0.84
Third quintile	1,887	3.00	0.82
Fourth quintile	2,135	3.18	0.80
Fifth quintile (highest)	2,434	3.37	0.81
School type			
Public school	8,483	3.00	0.86
Private school	2,175	3.16	0.80

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

As noted earlier, SRS score statistics for subpopulations, with scores for third graders shown separately from those of first and second graders, are presented in tables 6-22 through 6-27.

Table 6-22. Score breakdown, Teacher Social Rating Scale (SRS), approaches to learning, by third graders, first and second graders, and population subgroup: School year 2001–02

Characteristic	Third graders			First and second graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,566	3.03	0.69	989	2.72	0.69
Sex						
Male	5,195	2.87	0.69	616	2.63	0.68
Female	5,371	3.19	0.64	373	2.89	0.68
Race/ethnicity						
White, non-Hispanic	6,573	3.08	0.67	473	2.83	0.67
Black, non-Hispanic	1,136	2.82	0.71	240	2.49	0.68
Hispanic, race specified	799	3.03	0.68	91	2.80	0.74
Hispanic, race not specified	836	2.95	0.69	75	2.67	0.68
Asian	650	3.34	0.56	45	2.90	0.68
Hawaiian, other Pacific Islander	148	2.93	0.66	6	2.75	0.28
American Indian/Alaska Native	145	2.98	0.65	40	2.76	0.56
More than one race, non-Hispanic	266	2.99	0.66	19	2.71	0.63
Socioeconomic status						
First quintile (lowest)	1,190	2.79	0.72	246	2.57	0.69
Second quintile	1,625	2.95	0.68	189	2.63	0.60
Third quintile	1,890	3.00	0.67	150	2.79	0.58
Fourth quintile	2,180	3.15	0.65	112	2.98	0.73
Fifth quintile (highest)	2,536	3.25	0.63	111	3.10	0.64
School type						
Public school	8,272	3.02	0.69	883	2.70	0.69
Private school	2,291	3.14	0.64	106	3.08	0.61

[1] Standard deviation.
NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-23. Score breakdown, Teacher Social Rating Scale (SRS), self-control, by third graders, first and second graders, and population subgroup: School year 2001–02

Characteristic	Third graders			First and second graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,462	3.19	0.62	985	2.99	0.65
Sex						
Male	5,147	3.08	0.64	615	2.92	0.65
Female	5,315	3.30	0.57	370	.10	.65
Race/ethnicity						
White, non-Hispanic	6,509	3.24	0.60	469	3.11	0.60
Black, non-Hispanic	1,132	2.96	0.68	240	2.74	0.70
Hispanic, race specified	781	3.20	0.60	91	3.05	0.65
Hispanic, race not specified	832	3.17	0.61	75	2.96	0.62
Asian	640	3.40	0.54	45	3.12	0.67
Hawaiian, other Pacific Islander	148	3.20	0.61	6	2.93	0.58
American Indian/Alaska Native	145	3.13	0.57	40	2.89	0.60
More than one race, non-Hispanic	262	3.13	0.62	19	2.97	0.72
Socioeconomic status						
First quintile (lowest)	1,180	3.03	0.65	247	2.88	0.65
Second quintile	1,609	3.15	0.61	189	3.01	0.61
Third quintile	1,874	3.18	0.62	149	3.05	0.60
Fourth quintile	2,167	3.27	0.60	12	.20	.69
Fifth quintile (highest)	2,499	3.33	0.56	108	3.25	0.51
School type						
Public school	8,217	3.18	0.62	880	2.98	0.66
Private school	2,242	3.27	0.58	05	.16	.60

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-24. Score breakdown, Teacher Social Rating Scale (SRS), interpersonal, by third graders, first and second graders, and population subgroup: School year 2001–02

Characteristic	Third graders			First and second graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,439	3.08	0.65	975	2.85	0.69
Sex						
Male	5,116	2.95	0.66	608	2.78	0.69
Female	5,323	3.21	0.62	367	2.99	0.68
Race/ethnicity						
White, non-Hispanic	6,512	3.12	0.65	469	2.96	0.68
Black, non-Hispanic	1,123	2.89	0.69	232	2.67	0.72
Hispanic, race specified	783	3.07	0.64	91	2.95	0.70
Hispanic, race not specified	818	3.06	0.63	75	2.79	0.67
Asian	639	3.26	0.57	44	2.72	0.49
Hawaiian, other Pacific Islander	147	3.12	0.65	6	2.93	0.31
American Indian, Alaska Native	143	2.95	0.59	39	2.74	0.59
More than one race, non-Hispanic	261	2.98	0.64	19	2.76	0.51
Socioeconomic status						
First quintile (lowest)	1,172	2.92	0.67	241	2.70	0.70
Second quintile	1,601	3.01	0.65	185	2.90	0.67
Third quintile	1,872	3.07	0.64	149	2.90	0.61
Fourth quintile	2,152	3.14	0.65	110	3.08	0.67
Fifth quintile (highest)	2,511	3.24	0.62	111	3.13	0.63
School type						
Public school	8,171	3.06	0.66	870	2.83	0.69
Private school	2,266	3.19	0.61	105	3.15	0.61

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-25. Score breakdown, Teacher Social Rating Scale (SRS), externalizing problem behaviors, by third graders, first and second graders, and population subgroup: School year 2001–02

Characteristic	Third graders			First and second graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,543	1.70	0.61	987	1.94	0.71
Sex						
Male	5,188	1.84	0.65	615	2.04	0.72
Female	5,355	1.58	0.53	372	1.76	0.64
Race/ethnicity						
White, non-Hispanic	6,559	1.66	0.58	472	1.85	0.64
Black, non-Hispanic	1,132	1.97	0.70	241	2.13	0.82
Hispanic, race specified	797	1.69	0.58	90	1.91	0.72
Hispanic, race not specified	835	1.69	0.61	75	1.88	0.65
Asian	650	1.46	0.44	44	1.90	0.76
Hawaiian, other Pacific Islander	148	1.74	0.58	6	2.23	0.67
American Indian/Alaska Native	144	1.74	0.52	40	2.01	0.62
More than one race, non-Hispanic	265	1.73	0.63	19	1.75	0.56
Socioeconomic status						
First quintile (lowest)	1,188	1.83	0.67	246	2.03	0.74
Second quintile	1,620	1.72	0.60	189	1.90	0.72
Third quintile	1,888	1.75	0.61	149	1.83	0.62
Fourth quintile	2,176	1.67	0.60	112	1.89	0.70
Fifth quintile (highest)	2,528	1.57	0.53	111	1.75	0.65
School type						
Public school	8,259	1.72	0.61	882	1.95	0.71
Private school	2,281	1.63	0.55	105	1.74	0.56

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-26. Score breakdown, Teacher Social Rating Scale (SRS), internalizing problem behaviors, by third graders, first and second graders, and population subgroup: School year 2001–02

Characteristic	Third graders			First and second graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,455	1.65	0.55	978	1.81	0.61
Sex						
Male	5,139	1.67	0.56	608	1.81	0.62
Female	5,316	1.63	0.53	370	1.81	0.59
Race/ethnicity						
White, non-Hispanic	6,527	1.65	0.55	467	1.77	0.56
Black, non-Hispanic	1,100	1.65	0.54	237	1.92	0.71
Hispanic, race specified	790	1.65	0.56	91	1.72	0.57
Hispanic, race not specified	828	1.64	0.53	74	1.73	0.60
Asian	643	1.46	0.41	44	1.67	0.43
Hawaiian, other Pacific Islander	147	1.72	0.54	6	1.35	0.39
American Indian/Alaska Native	144	1.76	0.63	40	1.95	0.53
More than one race, non-Hispanic	263	1.69	0.57	19	1.95	0.41
Socioeconomic status						
First quintile (lowest)	1,171	1.78	0.61	243	1.9	0.63
Second quintile	1,603	1.70	0.57	188	1.83	0.56
Third quintile	1,877	1.62	0.52	148	1.74	0.59
Fourth quintile	2,163	1.60	0.51	111	1.71	0.57
Fifth quintile (highest)	2,512	1.54	0.49	111	1.66	0.54
School type						
Public school	8,186	1.66	0.55	876	1.82	0.61
Private school	2,266	1.57	0.50	102	1.68	0.46

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-27. Score breakdown, Teacher Social Rating Scale (SRS), peer relations: self-control + interpersonal, by third graders, first and second graders, and population subgroup: School year 2001–02

Characteristic	Third graders			First and second graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,527	3.13	0.61	988	2.91	0.63
Sex						
Male	5,169	3.01	0.62	616	2.84	0.63
Female	5,358	3.25	0.57	372	3.04	0.62
Race/ethnicity						
White, non-Hispanic	6,559	3.17	0.59	471	3.03	0.60
Black, non-Hispanic	1,134	2.92	0.65	241	2.70	0.67
Hispanic, race specified	790	3.13	0.59	91	3.00	0.63
Hispanic, race not specified	829	3.11	0.58	75	2.86	0.61
Asian	645	3.33	0.53	45	2.90	0.54
Hawaiian, other Pacific Islander	147	3.16	0.60	6	2.93	0.39
American Indian/Alaska Native	145	3.03	0.55	40	2.80	0.52
More than one race, non-Hispanic	265	3.05	0.60	19	2.85	0.58
Socioeconomic status						
First quintile (lowest)	1,186	2.97	0.63	247	2.78	0.63
Second quintile	1,612	3.08	0.61	189	2.95	0.60
SES: third quintile	1,883	3.12	0.60	150	2.97	0.57
Fourth quintile	2,177	3.20	0.59	112	3.13	0.65
Fifth quintile (highest)	2,529	3.28	0.56	110	3.18	0.54
School type						
Public school	8,238	3.12	0.61	882	2.90	0.64
Private school	2,286	3.23	0.56	106	3.15	0.57

[1] Standard deviation.
NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

As noted earlier, SDQ score statistics for subpopulations are presented in tables 6-28 through 6-33.

Table 6-28. Score breakdown, Self-Description Questionnaire (SDQ), perceived interest/ competence in reading, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,351	3.26	0.66
Sex			
Male	7,279	3.18	0.69
Female	7,072	3.35	0.62
Race/ethnicity			
White, non-Hispanic	8,120	3.24	0.67
Black, non-Hispanic	1,871	3.30	0.66
Hispanic, race specified	1,260	3.29	0.66
Hispanic, race not specified	1,325	3.31	0.62
Asian	956	3.28	0.61
Hawaiian, other Pacific Islander	170	3.29	0.57
American Indian/Alaska Native	250	3.31	0.65
More than one race, non-Hispanic	379	3.24	0.67
Socioeconomic status			
First quintile (lowest)	2,004	3.28	0.64
Second quintile	2,250	3.20	0.70
Third quintile	2,453	3.22	0.68
Fourth quintile	2,694	3.27	0.67
Fifth quintile (highest)	3,162	3.32	0.63
School type			
Public school	11,669	3.27	0.66
Private school	2,633	3.23	0.67

[1] Standard deviation.
NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-29. Score breakdown, Self-Description Questionnaire (SDQ), perceived interest/ competence in mathematics, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,351	3.16	0.79
Sex			
Male	7,279	3.24	0.76
Female	7,072	3.08	0.81
Race/ethnicity			
White, non-Hispanic	8,120	3.11	0.80
Black, non-Hispanic	1,871	3.24	0.80
Hispanic, race specified	1,260	3.22	0.73
Hispanic, race not specified	1,325	3.25	0.73
Asian	956	3.23	0.73
Hawaiian, other Pacific Islander	170	3.12	0.73
American Indian/Alaska Native	250	3.17	0.73
More than one race, non-Hispanic	379	3.21	0.77
Socioeconomic status			
First quintile (lowest)	2,004	3.23	0.76
Second quintile	2,250	3.14	0.82
Third quintile	2,453	3.16	0.80
Fourth quintile	2,694	3.14	0.80
Fifth quintile (highest)	3,162	3.12	0.77
School type			
Public school	11,669	3.19	0.78
Private school	2,633	2.98	0.84

[1] Standard deviation.
NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-30. Score breakdown, Self-Description Questionnaire (SDQ), perceived interest/ competence in peer relations, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,350	3.03	0.65
Sex			
Male	7,278	2.99	0.66
Female	7,072	3.08	0.64
Race/ethnicity			
White, non-Hispanic	8,119	3.02	0.63
Black, non-Hispanic	1,871	3.10	0.70
Hispanic, race specified	1,260	3.05	0.65
Hispanic, race not specified	1,325	3.02	0.65
Asian	956	2.87	0.63
Hawaiian, other Pacific Islander	170	2.97	0.68
American Indian/Alaska Native	250	2.94	0.70
More than one race, non-Hispanic	379	3.03	0.73
Socioeconomic status			
First quintile (lowest)	2,004	3.04	0.69
Second quintile	2,250	3.00	0.66
Third quintile	2,453	3.02	0.65
Fourth quintile	2,694	3.02	0.64
Fifth quintile (highest)	3,161	3.05	0.61
School type			
Public school	11,669	3.03	0.66
Private school	2,632	3.01	0.61

[1] Standard deviation.
NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-31. Score breakdown, Self-Description Questionnaire (SDQ), perceived interest/ competence in all subjects, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,351	2.92	0.66
Sex			
Male	7,279	2.89	0.66
Female	7,072	2.96	0.65
Race/ethnicity			
White, non-Hispanic	8,120	2.88	0.66
Black, non-Hispanic	1,871	2.99	0.69
Hispanic, race specified	1,260	2.97	0.63
Hispanic, race not specified	1,325	3.01	0.60
Asian	956	2.96	0.58
Hawaiian, other Pacific Islander	170	2.92	0.59
American Indian/Alaska Native	250	2.95	0.67
More than one race, non-Hispanic	379	2.97	0.66
Socioeconomic status			
First quintile (lowest)	2,004	2.99	0.65
Second quintile	2,250	2.89	0.69
Third quintile	2,453	2.90	0.67
Fourth quintile	2,694	2.91	0.64
Fifth quintile (highest)	3,162	2.91	0.62
School type			
Public school	11,669	2.94	0.65
Private school	2,633	2.80	0.68

[1] Standard deviation.
NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-32. Score breakdown, Self-Description Questionnaire (SDQ), internalizing problems, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,351	2.23	0.74
Sex			
Male	7,279	2.22	0.73
Female	7,072	2.24	0.74
Race/ethnicity			
White, non-Hispanic	8,120	2.08	0.70
Black, non-Hispanic	1,871	2.49	0.77
Hispanic, race specified	1,260	2.37	0.73
Hispanic, race not specified	1,325	2.49	0.72
Asian	956	2.13	0.66
Hawaiian, other Pacific Islander	170	2.47	0.68
American Indian/Alaska Native	250	2.39	0.75
More than one race, non-Hispanic	379	2.11	0.75
Socioeconomic status			
First quintile (lowest)	2,004	2.57	0.72
Second quintile	2,250	2.33	0.75
Third quintile	2,453	2.19	0.71
Fourth quintile	2,694	2.06	0.68
Fifth quintile (highest)	3,162	1.89	0.63
School type			
Public school	11,669	2.25	0.75
Private school	2,633	2.04	0.66

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

Table 6-33. Score breakdown, Self-Description Questionnaire (SDQ), externalizing problems, by population subgroup, spring-third grade: School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,351	2.02	0.71
Sex			
Male	7,279	2.12	0.72
Female	7,072	1.92	0.68
Race/ethnicity			
White, non-Hispanic	8,120	1.93	0.67
Black, non-Hispanic	1,871	2.27	0.78
Hispanic, race specified	1,260	2.07	0.72
Hispanic, race not specified	1,325	2.11	0.69
Asian	956	1.85	0.60
Hawaiian, other Pacific Islander	170	2.34	0.72
American Indian/Alaska Native	250	2.22	0.75
More than one race, non-Hispanic	379	2.01	0.69
Socioeconomic status			
First quintile (lowest)	2,004	2.28	0.75
Second quintile	2,250	2.10	0.72
Third quintile	2,453	2.01	0.69
Fourth quintile	2,694	1.89	0.64
Fifth quintile (highest)	3,162	1.78	0.59
School type			
Public school	11,669	2.04	0.71
Private school	2,633	1.89	0.64

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first grade variable for comparison. 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 2002.

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. (2001). *Measures of socio-emotional development in middle childhood* (NCES 2001-03). National Center for Education Statistics Working Paper. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement.

Atkins-Burnett, S., Meisels, S. J., and Correnti, R. (2000). Analysis to develop third-grade indirect cognitive assessments and socioemotional measures. In *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) Spring 2000 Field Test Report.* (Prepared under contract to the U.S. Department of Education, National Center for Education Statistics.) Rockville, MD: Westat.

Bryk, A.S., and Raudenbush, S.W. (1992). *Hierarchical linear models, applications and data analysis methods.* Newbury Park, CA: Sage Publications.

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 multitrait-multimethod 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.

Gresham, F., and Elliot, S. (1990). *Social skills rating system*. Circle Pines, MN: American Guidance Services, Inc.

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.

Holt (1986). *Science*. New York: Author.

Kirsch, I.S., et al. (1993). Adult literacy in America: A first look at the results of the National Adult Literacy Survey. Washington, DC: National Center for Education Statistics.

Linacre, J.M., and Wright, B.D. (2000). *A user's guide to Winsteps Ministep Rasch model computer programs*. Chicago, IL: MESA 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. (1990). *Self-Description Questionnaire manual.* Campbelltown, N.S.W, Australia: University of Western Sydney, Macarthur.

Marsh, H.W., Chessor, D., Craven, R., & 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.

Meisels, S.J., and Perry, N.E. (1996). *How accurate are teacher judgments of student's academic performance?* (Working Paper No. 96-08). Washington, DC: National Center for Education Statistics.

Meisels, S.J., Marsden, D.B., Wiske, M.S., and Henderson, L.W. (1997*). The Early Screening Inventory – Revised.* Ann Arbor, MI: Rebus, Inc.

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., et al. (1992). Scaling procedures in NAEP. *Journal of Education 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.

National Academy of Sciences. (1995). *National science education standards.* Washington, DC: Author.

National Assessment Governing Board (NAGB). (1994b). *Geography Framework for the 1994 National Assessment of Educational Progress*. Washington, DC: Government Printing Office.

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

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

National Council of Teachers of Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: Author.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institute.

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., et. al. (1985). *Psychometric analysis of the NLS-72 and the High School And Beyond test batteries*. (NCES Report No.85-217). Washington, DC: National Center for Education Statistics.

Rock, D.A., et. al. (1995). *Psychometric report for the NELS:88 base year test battery.* (NCES Report No.95-382). Washington, DC: National Center for Education Statistics.

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

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

Smith, R.M., Schumacker, R.E., and Bush, M.J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement,* 2(1), 66-78.

U.S. Department of Education, National Center for Education Statistics (2004). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), User's Manual for the ECLS-K Public-Use Data File and Electronic Code Book* (NCES 2004-001). Washington, DC: Author.

U.S. Department of Education, National Center for Education Statistics (2002). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for Kindergarten Through the First Grade* (NCES 2002-05), by Donald A. Rock and Judith M. Pollack, Educational Testing Service, Elvira Germino Hausken, project officer. Washington, DC..

U.S. Department of Education, National Center for Education Statistics. (2001). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) First Grade Public-Use User's Manual* (NCES 2001-148). Washington, DC: Author.

U.S. Department of Education, National Center for Education Statistics (2003). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), User's Manual for the ECLS-K Third Grade Restricted-Use Data File and Electronic Code Book* (NCES 2003-003). Washington, DC: Author.

Woodcock, R.W., McGrew, K.S., and Werder, J.K. (1996). *Woodcock-McGrew-Werder Mini-Battery of Achievement.* Itasca, IL: Riverside Publishing.

Wright, B.D. (1999). Fundamental measurement for psychology. In S. Embretson and S. L. Hershberger (Eds.) *The new rules of measurement: What every psychologist and educator should know* (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

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.

This page is intentionally left blank.

APPENDIX A

SCORE STATISTICS FOR DIRECT COGNITIVE MEASURES FOR SELECTED SUBGROUPS

Table A1. Reading routing test number right, third grade assessment (range of possible values: 0 to 15): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,246	9.94	2.75
Sex			
Male	7,204	9.81	2.87
Female	7,042	10.08	2.61
Race/ethnicity			
White, non-Hispanic	8,082	10.54	2.52
Black, non-Hispanic	1,840	8.84	2.73
Hispanic, race specified	1,252	9.51	2.82
Hispanic, race not specified	1,314	8.82	2.94
Asian	956	10.26	2.43
Hawaiian, other Pacific Islander	171	9.52	2.65
American Indian/Alaska Native	232	8.08	2.99
More than one race, non-Hispanic	379	10.11	2.75
Socioeconomic status			
First quintile	1,964	8.22	2.80
Second quintile	2,230	9.45	2.63
Third quintile	2,437	10.10	2.47
Fourth quintile	2,688	10.72	2.31
Fifth quintile	3,158	11.56	2.15
School type			
Public school	11,575	9.79	2.77
Private school	2,623	11.11	2.26

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten-first 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 2002.

Table A2. Mathematics routing test number right, third grade assessment (range of possible values: 0 to 17): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,349	8.79	4.39
Sex			
Male	7,277	9.07	4.51
Female	7,072	8.50	4.25
Race/ethnicity			
White, non-Hispanic	8,116	9.92	4.14
Black, non-Hispanic	1,871	6.45	4.11
Hispanic, race specified	1,260	7.75	4.22
Hispanic, race not specified	1,324	7.17	4.07
Asian	956	9.79	4.46
Hawaiian, other Pacific Islander	172	7.51	3.94
American Indian/Alaska Native	250	6.25	3.92
More than one race, non-Hispanic	380	9.02	4.42
Socioeconomic status			
First quintile	2,001	6.14	3.98
Second quintile	2,250	7.90	4.18
Third quintile	2,452	8.91	4.03
Fourth quintile	2,693	10.13	3.99
Fifth quintile	3,163	11.53	3.76
School type			
Public school	11,670	8.66	4.41
Private school	2,631	9.92	4.09

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first 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 2002.

Table A3. Science routing test number right, third grade assessment (range of possible values: 0 to 15): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,339	8.25	3.38
Sex			
Male	7,267	8.57	3.41
Female	7,072	7.90	3.31
Race/ethnicity			
White, non-Hispanic	8,110	9.42	2.99
Black, non-Hispanic	1,869	6.09	2.97
Hispanic, race specified	1,259	7.01	3.28
Hispanic, race not specified	1,325	6.33	3.16
Asian	956	8.33	3.48
Hawaiian, other Pacific Islander	172	7.44	3.20
American Indian/Alaska Native	249	6.74	3.15
More than one race, non-Hispanic	379	8.64	3.13
Socioeconomic status			
First quintile	1,999	5.84	2.99
Second quintile	2,249	7.69	3.09
Third quintile	2,452	8.53	3.02
Fourth quintile	2,692	9.41	2.95
Fifth quintile	3,162	10.51	2.77
School type			
Public school	11,657	8.10	3.37
Private school	2,633	9.37	3.17

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first 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 2002.

Table A4. Reading IRT scale score, K–3 scale (range of possible values: 0 to 154): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	26.9	9.7	18,935	37.9	13.0	5,053	44.3	16.5	16,336	66.6	20.8	14,246	106.1	20.7
Sex															
Male	8,984	26.3	9.8	9,688	36.9	13.0	2,556	43.0	16.3	8,349	64.8	21.1	7,204	104.2	21.3
Female	8,640	27.6	9.6	9,247	39.0	13.0	2,497	45.8	16.5	7,987	68.5	20.2	7,042	108.1	19.9
Race/ethnicity															
White, non-Hispanic	10,433	28.2	9.9	11,073	39.7	13.3	2,935	46.8	17.0	9,435	70.4	20.6	8,082	111.6	19.0
Black, non-Hispanic	2,854	24.5	7.8	2,968	34.3	11.1	782	40.1	13.5	2,371	59.5	19.0	1,840	97.0	19.7
Hispanic, race specified	1,182	24.9	9.2	1,315	36.1	12.0	322	42.4	14.3	1,233	63.0	19.5	1,252	101.4	20.7
Hispanic, race not specified	1,195	23.3	7.4	1,423	33.6	10.8	377	37.6	12.8	1,335	58.4	18.3	1,314	95.0	20.7
Asian	896	31.5	14.0	1,088	43.9	16.9	257	52.6	22.6	1,042	73.3	22.2	956	108.8	18.6
Hawaiian, other Pacific Islander	186	25.9	9.6	202	35.3	11.3	93	38.0	12.9	188	63.2	18.9	171	101.3	18.8
American Indian/Alaska Native	354	21.4	6.5	344	31.6	9.7	126	32.1	10.3	298	52.6	17.4	232	90.2	21.3
More than one race, non-Hispanic	476	26.9	11.1	473	37.6	14.0	152	44.2	15.4	397	67.6	20.9	379	106.8	19.9
Socioeconomic status															
First quintile	2,594	21.8	5.9	2,917	31.3	9.0	753	35.4	11.0	2,363	54.6	17.1	1,964	91.6	20.2
Second quintile	3,271	24.6	7.6	3,503	35.1	11.1	925	40.2	13.3	2,796	62.9	18.9	2,230	102.1	19.6
Third quintile	3,470	26.2	8.0	3,686	37.4	11.2	997	44.7	15.1	3,003	67.0	18.7	2,437	107.4	18.3
Fourth quintile	3,650	28.6	9.6	3,909	40.4	12.6	1,019	47.5	16.0	3,173	71.2	19.2	2,688	112.5	17.8
Fifth quintile	3,880	32.8	12.3	4,152	45.3	15.8	1,159	53.2	19.5	3,642	78.6	21.1	3,158	119.8	15.5
School type															
Public school	13,736	26.1	9.2	14,578	36.9	12.4	3,809	43.4	16.0	12,998	65.2	20.3	11,575	104.9	20.9
Private school	3,888	31.3	11.5	4,357	43.5	15.0	1,042	51.3	17.4	3,279	76.0	20.6	2,623	114.8	17.1

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A5. Mathematics IRT scale score, K–3 scale (range of possible values: 0 to 123): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	21.0	8.7	19,647	30.8	11.3	5,226	37.5	13.4	16,641	53.7	16.1	14,349	83.2	18.3
Sex															
Male	9,479	21.0	9.2	10,041	30.9	11.9	2,644	37.6	14.2	8,506	54.2	17.0	7,277	84.6	18.8
Female	9,156	20.9	8.1	9,606	30.6	10.7	2,582	37.3	12.4	8,135	53.1	15.0	7,072	81.8	17.6
Race/ethnicity															
White, non-Hispanic	10,433	23.2	8.9	11,071	33.7	11.4	2,935	40.8	13.4	9,436	57.8	16.0	8,116	88.1	16.8
Black, non-Hispanic	2,855	17.9	6.5	2,962	26.2	9.2	781	32.6	11.5	2,371	45.8	13.5	1,871	73.0	17.3
Hispanic, race specified	1,588	18.1	7.3	1,624	27.4	10.1	389	34.7	11.9	1,354	50.1	15.2	1,260	79.1	18.1
Hispanic, race not specified	1,800	16.2	6.4	1,834	25.1	9.3	486	30.4	10.8	1,518	46.9	13.1	1,324	76.2	17.5
Asian	897	24.7	10.4	1,088	34.6	12.8	256	41.6	15.6	1,042	56.1	17.1	956	87.7	18.7
Hawaiian, other Pacific Islander	187	19.2	7.5	202	27.4	9.7	93	31.8	9.7	188	46.8	12.5	172	78.7	16.5
American Indian/Alaska Native	354	16.6	6.8	345	25.9	9.3	126	28.1	11.0	298	45.6	13.2	250	72.6	17.1
More than one race, non-Hispanic	473	20.9	8.7	472	30.3	10.7	151	36.2	11.9	397	54.0	15.9	380	84.4	17.9
Socioeconomic status															
First quintile	3,269	15.7	5.8	3,426	24.0	8.5	895	29.2	10.9	2,572	44.7	13.5	2,001	71.4	17.1
Second quintile	3,429	18.8	6.9	3,607	28.5	9.8	942	34.3	11.6	2,839	50.2	14.6	2,250	79.4	17.1
Third quintile	3,546	20.9	7.4	3,721	30.9	10.0	1,001	38.0	11.3	3,017	54.0	14.5	2,452	84.0	16.3
Fourth quintile	3,676	23.1	8.3	3,921	33.5	10.7	1,023	40.4	11.8	3,178	57.6	14.7	2,693	89.2	16.2
Fifth quintile	3,893	26.9	10.1	4,161	38.0	12.4	1,158	46.4	14.5	3,644	63.9	15.9	3,163	95.1	14.6
School type															
Public school	14,701	20.2	8.3	15,259	29.9	11.0	3,971	36.7	13.2	13,292	52.8	16.0	11,670	82.7	18.4
Private school	3,934	25.2	9.7	4,388	35.7	12.0	1,043	43.8	12.9	3,286	60.0	15.1	2,631	88.2	16.3

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A6. Science IRT scale score, K–3 scale (range of possible values: 0 to 62): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,339	33.46	10.01
Sex			
Male	7,267	34.52	10.11
Female	7,072	32.33	9.77
Race/ethnicity			
White, non-Hispanic	8,110	37.12	8.84
Black, non-Hispanic	1,869	26.82	8.55
Hispanic, race specified	1,259	29.52	9.59
Hispanic, race not specified	1,325	27.40	8.92
Asian	956	33.63	10.25
Hawaiian, other Pacific Islander	172	30.55	9.06
American Indian/Alaska Native	249	28.52	9.03
More than one race, non-Hispanic	379	34.97	9.06
Socioeconomic status			
First quintile	1,999	25.93	8.62
Second quintile	2,249	31.67	8.98
Third quintile	2,452	34.30	8.61
Fourth quintile	2,692	37.09	8.59
Fifth quintile	3,162	40.67	8.24
School type			
Public school	11,657	32.99	9.99
Private school	2,633	37.02	9.28

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first 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 2002.

Table A7. Reading T-scores, standardized within round (range of possible values: 0 to 96): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	50.0	10.0	18,935	50.0	10.0	5,053	50.0	10.0	16,336	50.0	10.0	14,246	50.0	10.0
Sex															
Male	8,984	49.2	10.0	9,688	49.1	10.2	2,556	49.0	10.2	8,349	49.0	10.4	7,204	49.1	10.3
Female	8,640	50.8	9.9	9,247	51.0	9.7	2,497	51.1	9.7	7,987	51.0	9.4	7,042	51.0	9.6
Race/ethnicity															
White, non-Hispanic	10,433	51.7	9.7	11,073	51.6	9.5	2,935	51.7	9.5	9,435	51.9	9.3	8,082	52.6	9.3
Black, non-Hispanic	2,854	47.3	9.2	2,968	46.9	9.8	782	47.3	9.5	2,371	46.5	10.5	1,840	45.6	9.4
Hispanic, race specified	1,182	47.5	9.9	1,315	48.5	10.0	322	49.0	9.3	1,233	48.4	9.8	1,252	47.7	9.9
Hispanic, race not specified	1,195	45.6	9.2	1,423	46.3	10.0	377	45.1	10.1	1,335	46.2	9.8	1,314	44.7	9.9
Asian	896	54.2	11.6	1,088	54.2	10.5	257	54.1	11.5	1,042	52.8	10.1	956	51.3	8.9
Hawaiian, other Pacific Islander	186	48.5	10.6	202	47.8	10.0	93	45.5	9.2	188	48.8	8.7	171	47.8	8.8
American Indian/Alaska Native	354	43.0	8.8	344	44.3	9.7	126	40.3	9.8	298	42.9	10.1	232	42.5	10.4
More than one race, non-Hispanic	476	49.7	10.6	473	49.6	10.2	152	50.0	9.8	397	50.5	10.1	379	50.3	9.6
Socioeconomic status															
First quintile	2,594	43.8	7.9	2,917	44.2	9.1	753	43.6	9.4	2,363	44.0	10.2	1,964	43.1	9.8
Second quintile	3,271	47.5	8.7	3,503	47.8	9.5	925	47.4	9.3	2,796	48.4	9.7	2,230	48.0	9.3
Third quintile	3,470	49.5	8.9	3,686	50.0	9.1	997	50.7	9.0	3,003	50.5	8.8	2,437	50.6	8.6
Fourth quintile	3,650	52.2	9.4	3,909	52.4	8.9	1,019	52.5	8.6	3,173	52.4	8.4	2,688	53.1	8.6
Fifth quintile	3,880	56.3	10.2	4,152	55.6	9.4	1,159	55.4	9.2	3,642	55.4	8.4	3,158	56.7	8.0
School type															
Public school	13,736	49.1	9.7	14,578	49.2	9.9	3,809	49.4	9.9	12,998	49.4	10.0	11,575	49.5	10.1
Private school	3,888	54.8	10.0	4,357	54.3	9.7	1,042	54.6	8.7	3,279	54.3	8.6	2,623	54.2	8.5

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A8. Mathematics T-scores, standardized within round (range of possible values: 0 to 96): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	50	10	19,647	50.0	10.0	5,226	50.0	10.0	16,641	50.0	10.0	14,349	50.0	10.0
Sex															
Male	9,479	49.9	10.4	10,041	50.0	10.3	2,644	49.9	10.6	8,506	50.2	10.5	7,277	50.8	10.4
Female	9,156	50.1	9.5	9,606	50.0	9.6	2,582	50.1	9.4	8,135	49.8	9.4	7,072	49.2	9.5
Race/ethnicity															
White, non-Hispanic	10,433	52.7	9.5	11,071	52.7	9.3	2,935	52.6	9.2	9,436	52.5	9.4	8,116	52.6	9.4
Black, non-Hispanic	2,855	46.5	8.7	2,962	45.9	9.2	781	46.3	9.8	2,371	45.2	9.9	1,871	44.5	9.2
Hispanic, race specified	1,588	46.4	9.6	1,624	46.9	9.8	389	48.0	9.6	1,354	47.9	10.0	1,260	47.8	9.8
Hispanic, race not specified	1,800	43.8	9.2	1,834	44.6	9.7	486	44.4	9.7	1,518	46.1	9.1	1,324	46.1	9.3
Asian	897	54.1	10.1	1,088	53.2	9.8	256	52.8	10.3	1,042	51.4	10.0	956	52.6	10.4
Hawaiian, other Pacific Islander	187	48.1	9.3	202	47.1	9.1	93	46.1	7.9	188	46.2	8.4	172	47.4	8.8
American Indian/Alaska Native	354	44.2	9.6	345	45.6	9.3	126	42.1	10.5	298	45.2	9.3	250	44.4	9.0
More than one race, non-Hispanic	473	50.1	9.4	472	49.8	9.3	151	49.2	9.4	397	50.1	10.0	380	50.5	9.8
Socioeconomic status															
First quintile	3,269	43.2	8.7	3,426	43.6	9.1	895	43.3	9.7	2,572	44.4	10.0	2,001	43.6	9.2
Second quintile	3,429	47.6	9	3,607	48.1	9.4	942	47.8	9.4	2,839	48.0	9.8	2,250	47.9	9.1
Third quintile	3,546	50.3	8.7	3,721	50.5	8.8	1,001	50.8	8.6	3,017	50.5	8.9	2,452	50.3	8.7
Fourth quintile	3,676	52.8	8.9	3,921	52.7	8.8	1,023	52.6	8.2	3,178	52.6	8.5	2,693	53.2	9.1
Fifth quintile	3,893	56.6	9.4	4,161	56.1	9.1	1,158	56.2	8.9	3,644	55.8	8.3	3,163	56.7	8.7
School type															
Public school	14,701	49.2	9.8	15,259	49.2	9.9	3,971	49.4	10.1	13,292	49.5	10.1	11,670	49.7	10.1
Private school	3,934	54.9	9.4	4,388	54.3	9.3	1,043	54.8	8.0	3,286	53.8	8.3	2,631	52.7	9.0

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A9. Science T-scores, standardized within round (range of possible values: 0 to 96): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,339	50.00	10.00
Sex			
Male	7,267	51.07	10.08
Female	7,072	48.86	9.79
Race/ethnicity			
White, non-Hispanic	8,110	53.64	8.71
Black, non-Hispanic	1,869	43.40	8.86
Hispanic, race specified	1,259	46.09	9.71
Hispanic, race not specified	1,325	43.97	9.18
Asian	956	50.20	10.08
Hawaiian, other Pacific Islander	172	47.02	9.32
American Indian/Alaska Native	249	45.19	9.11
More than one race, non-Hispanic	379	51.53	8.99
Socioeconomic status			
First quintile	1,999	42.47	9.00
Second quintile	2,249	48.29	8.94
Third quintile	2,452	50.89	8.36
Fourth quintile	2,692	53.59	8.44
Fifth quintile	3,162	57.14	8.23
School type			
Public school	11,657	49.53	10.00
Private school	2,633	53.57	9.15

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first 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 2002.

Table A10. Reading IRT theta score, K–3 scale (range of possible values: -5 to 5): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	-1.11	0.57	18,935	-0.48	0.56	5,053	-0.22	0.57	16,336	0.45	0.53	14,246	1.26	0.39
Sex															
Male	8,984	-1.15	0.57	9,688	-0.53	0.57	2,556	-0.28	0.58	8,349	0.40	0.55	7,204	1.23	0.41
Female	8,640	-1.06	0.56	9,247	-0.42	0.54	2,497	-0.16	0.55	7,987	0.50	0.50	7,042	1.30	0.38
Race/ethnicity															
White, non-Hispanic	10,433	-1.01	0.55	11,073	-0.39	0.53	2,935	-0.12	0.54	9,435	0.55	0.49	8,082	1.37	0.37
Black, non-Hispanic	2,854	-1.26	0.52	2,968	-0.66	0.55	782	-0.37	0.54	2,371	0.26	0.55	1,840	1.09	0.37
Hispanic, race specified	1,182	-1.25	0.56	1,315	-0.56	0.56	322	-0.28	0.53	1,233	0.37	0.52	1,252	1.17	0.39
Hispanic, race not specified	1,195	-1.36	0.52	1,423	-0.69	0.56	377	-0.50	0.58	1,335	0.25	0.52	1,314	1.05	0.39
Asian	896	-0.87	0.66	1,088	-0.25	0.58	257	0.01	0.66	1,042	0.60	0.53	956	1.31	0.35
Hawaiian, other Pacific Islander	186	-1.19	0.60	202	-0.60	0.56	93	-0.47	0.53	188	0.39	0.46	171	1.17	0.35
American Indian/Alaska Native	354	-1.51	0.50	344	-0.80	0.54	126	-0.77	0.56	298	0.07	0.53	232	0.97	0.41
More than one race, non-Hispanic	476	-1.13	0.60	473	-0.50	0.57	152	-0.22	0.56	397	0.47	0.53	379	1.28	0.38
Socioeconomic status															
First quintile	2,594	-1.46	0.45	2,917	-0.80	0.51	753	-0.59	0.54	2,363	0.13	0.54	1,964	0.99	0.38
Second quintile	3,271	-1.25	0.49	3,503	-0.60	0.53	925	-0.37	0.53	2,796	0.37	0.51	2,230	1.19	0.37
Third quintile	3,470	-1.14	0.51	3,686	-0.48	0.51	997	-0.18	0.52	3,003	0.48	0.46	2,437	1.28	0.34
Fourth quintile	3,650	-0.99	0.53	3,909	-0.35	0.50	1,019	-0.08	0.49	3,173	0.58	0.44	2,688	1.38	0.34
Fifth quintile	3,880	-0.75	0.58	4,152	-0.17	0.52	1,159	0.09	0.53	3,642	0.73	0.44	3,158	1.53	0.32
School type															
Public school	13,736	-1.16	0.55	14,578	-0.52	0.55	3,809	-0.25	0.57	12,998	0.42	0.53	11,575	1.24	0.40
Private school	3,888	-0.84	0.57	4,357	-0.24	0.54	1,042	0.05	0.50	3,279	0.68	0.46	2,623	1.43	0.33

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A11. Mathematics IRT theta score, K–3 scale (range of possible values: -5 to 5): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	-1.04	0.58	19,647	-0.46	0.57	5,226	-0.15	0.58	16,641	0.45	0.53	14,349	1.27	0.48
Sex															
Male	9,479	-1.05	0.61	10,041	-0.46	0.59	2,644	-0.15	0.62	8,506	0.46	0.56	7,277	1.31	0.50
Female	9,156	-1.04	0.55	9,606	-0.46	0.54	2,582	-0.14	0.55	8,135	0.44	0.50	7,072	1.23	0.45
Race/ethnicity															
White, non-Hispanic	10,433	-0.88	0.55	11,071	-0.30	0.53	2,935	0.00	0.54	9,436	0.59	0.50	8,116	1.40	0.45
Black, non-Hispanic	2,855	-1.25	0.51	2,962	-0.69	0.52	781	-0.37	0.57	2,371	0.20	0.53	1,871	1.01	0.44
Hispanic, race specified	1,588	-1.25	0.56	1,624	-0.63	0.55	389	-0.27	0.56	1,354	0.34	0.53	1,260	1.17	0.47
Hispanic, race not specified	1,800	-1.40	0.53	1,834	-0.76	0.55	486	-0.48	0.57	1,518	0.24	0.49	1,324	1.09	0.45
Asian	897	-0.80	0.59	1,088	-0.27	0.56	256	0.01	0.60	1,042	0.53	0.53	956	1.40	0.50
Hawaiian, other Pacific Islander	187	-1.15	0.54	202	-0.62	0.52	93	-0.37	0.46	188	0.25	0.45	172	1.15	0.42
American Indian/Alaska Native	354	-1.38	0.55	345	-0.71	0.53	126	-0.61	0.61	298	0.20	0.50	250	1.00	0.43
More than one race, non-Hispanic	473	-1.04	0.55	472	-0.47	0.53	151	-0.20	0.55	397	0.46	0.53	380	1.30	0.47
Socioeconomic status															
First quintile	3,269	-1.43	0.50	3,426	-0.82	0.52	895	-0.54	0.57	2,572	0.15	0.53	2,001	0.96	0.44
Second quintile	3,429	-1.18	0.52	3,607	-0.56	0.53	942	-0.28	0.55	2,839	0.35	0.52	2,250	1.17	0.43
Third quintile	3,546	-1.02	0.50	3,721	-0.43	0.50	1,001	-0.10	0.50	3,017	0.48	0.48	2,452	1.28	0.42
Fourth quintile	3,676	-0.88	0.51	3,921	-0.30	0.50	1,023	0.00	0.48	3,178	0.59	0.45	2,693	1.42	0.44
Fifth quintile	3,893	-0.66	0.54	4,161	-0.11	0.52	1,158	0.21	0.52	3,644	0.76	0.44	3,163	1.59	0.42
School type															
Public school	14,701	-1.09	0.57	15,259	-0.50	0.56	3,971	-0.18	0.59	13,292	0.42	0.54	11,670	1.26	0.48
Private school	3,934	-0.76	0.55	4,388	-0.21	0.53	1,043	0.13	0.47	3,286	0.66	0.44	2,631	1.40	0.43

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A12. Science IRT theta score, K–3 scale (range of possible values: -5 to 5): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,339	-0.38	0.68
Sex			
Male	7,267	-0.31	0.68
Female	7,072	-0.46	0.66
Race/ethnicity			
White, non-Hispanic	8,110	-0.14	0.59
Black, non-Hispanic	1,869	-0.83	0.60
Hispanic, race specified	1,259	-0.65	0.66
Hispanic, race not specified	1,325	-0.79	0.62
Asian	956	-0.37	0.68
Hawaiian, other Pacific Islander	172	-0.59	0.63
American Indian/Alaska Native	249	-0.71	0.62
More than one race, non-Hispanic	379	-0.28	0.61
Socioeconomic status			
First quintile	1,999	-0.89	0.61
Second quintile	2,249	-0.50	0.60
Third quintile	2,452	-0.32	0.57
Fourth quintile	2,692	-0.14	0.57
Fifth quintile	3,162	0.10	0.56
School type			
Public school	11,657	-0.42	0.68
Private school	2,633	-0.14	0.62

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first 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 2002.

Table A13. Reading decoding score, third grade assessment (range of possible values: 0 to 4): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,228	1.06	1.24
Sex			
Male	7,198	1.03	1.24
Female	7,030	1.10	1.24
Race/ethnicity			
White, non-Hispanic	8,074	1.22	1.31
Black, non-Hispanic	1,836	0.63	0.99
Hispanic, race specified	1,251	1.04	1.15
Hispanic, race not specified	1,311	0.87	1.09
Asian	955	1.24	1.27
Hawaiian, other Pacific Islander	170	0.99	1.11
American Indian/Alaska Native	232	0.52	0.91
More than one race, non-Hispanic	379	1.16	1.33
Socioeconomic status			
First quintile	1,959	0.61	0.95
Second quintile	2,228	0.84	1.11
Third quintile	2,437	1.03	1.20
Fourth quintile	2,684	1.25	1.30
Fifth quintile	3,155	1.69	1.38
School type			
Public school	11,560	1.00	1.22
Private school	2620	1.52	1.35

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first 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 2002.

Table A14. Science: life science cluster score, third grade assessment (range of possible values: 0 to 5): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,272	2.98	1.43
Sex			
Male	7,240	3.13	1.39
Female	7,032	2.81	1.44
Race/ethnicity			
White, non-Hispanic	8,077	3.43	1.24
Black, non-Hispanic	1,861	2.16	1.36
Hispanic, race specified	1,250	2.50	1.44
Hispanic, race not specified	1,315	2.22	1.43
Asian	950	2.93	1.50
Hawaiian, other Pacific Islander	172	2.71	1.34
American Indian/Alaska Native	249	2.43	1.37
More than one race, non-Hispanic	378	3.18	1.34
Socioeconomic status			
First quintile	1,982	2.07	1.37
Second quintile	2,243	2.81	1.37
Third quintile	2,440	3.10	1.30
Fourth quintile	2,684	3.39	1.26
Fifth quintile	3,153	3.80	1.15
School type			
Public school	11,600	2.93	1.43
Private school	2,623	3.37	1.31

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first 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 2002.

Table A15. Science: earth science cluster score, third grade assessment (range of possible values: 0 to 5): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,298	2.69	1.37
Sex			
Male	7,245	2.82	1.38
Female	7,053	2.55	1.34
Race/ethnicity			
White, non-Hispanic	8,095	3.07	1.26
Black, non-Hispanic	1,859	2.00	1.31
Hispanic, race specified	1,256	2.26	1.37
Hispanic, race not specified	1,320	2.06	1.28
Asian	949	2.65	1.36
Hawaiian, other Pacific Islander	172	2.41	1.37
American Indian/Alaska Native	249	2.32	1.38
More than one race, non-Hispanic	378	2.85	1.26
Socioeconomic status			
First quintile	1,982	1.92	1.29
Second quintile	2,246	2.50	1.32
Third quintile	2,444	2.80	1.29
Fourth quintile	2,689	3.10	1.23
Fifth quintile	3,161	3.36	1.19
School type			
Public school	11,619	2.64	1.36
Private school	2,630	3.08	1.33

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first 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 2002.

Table A16. Science: physical science cluster score third grade assessment (range of possible values: 0 to 5): School year 2001–02

Characteristic	Round 5		
	Number	Mean	SD[1]
Total sample	14,245	2.60	1.33
Sex			
Male	7,219	2.64	1.35
Female	7,026	2.56	1.31
Race/ethnicity			
White, non-Hispanic	8,064	2.94	1.30
Black, non-Hispanic	1,849	1.95	1.16
Hispanic, race specified	1,245	2.29	1.24
Hispanic, race not specified	1,319	2.06	1.24
Asian	950	2.78	1.35
Hawaiian, other Pacific Islander	171	2.33	1.23
American Indian/Alaska Native	249	1.99	1.16
More than one race, non-Hispanic	378	2.63	1.33
Socioeconomic status			
First quintile	1,974	1.88	1.17
Second quintile	2,239	2.39	1.24
Third quintile	2,440	2.66	1.27
Fourth quintile	2,679	2.94	1.27
Fifth quintile	3,149	3.36	1.23
School type			
Public school	11,574	2.56	1.33
Private school	2,622	2.94	1.33

[1] Standard deviation.

NOTE: Table estimates are based on C5CW0 weight. There is no kindergarten/first 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 2002.

Table A17. Probability of proficiency, reading level 1: letter recognition (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.68	0.37	18,935	0.93	0.19	5,053	0.97	0.14	16,336	0.99	0.05	14,246	1.00	0.00
Sex															
Male	8,984	0.65	0.38	9,688	0.92	0.20	2,556	0.96	0.15	8,349	0.99	0.06	7,204	1.00	0.00
Female	8,640	0.71	0.36	9,247	0.95	0.17	2,497	0.97	0.12	7,987	1.00	0.04	7,042	1.00	0.00
Race/ethnicity															
White, non-Hispanic	10,433	0.74	0.34	11,073	0.96	0.15	2,935	0.98	0.11	9,435	1.00	0.04	8,082	1.00	0.00
Black, non-Hispanic	2,854	0.60	0.38	2,968	0.90	0.22	782	0.96	0.16	2,371	0.99	0.07	1,840	1.00	0.00
Hispanic, race specified	1,182	0.58	0.39	1,315	0.92	0.21	322	0.98	0.11	1,233	0.99	0.06	1,252	1.00	0.00
Hispanic, race not specified	1,195	0.52	0.40	1,423	0.87	0.26	377	0.92	0.22	1,335	0.99	0.06	1,314	1.00	0.00
Asian	896	0.80	0.31	1,088	0.97	0.11	257	0.99	0.08	1,042	1.00	0.04	956	1.00	0.00
Hawaiian, other Pacific Islander	186	0.62	0.39	202	0.91	0.21	93	0.96	0.11	188	1.00	0.00	171	1.00	0.00
American Indian/Alaska Native	354	0.40	0.39	344	0.85	0.27	126	0.84	0.29	298	0.99	0.05	232	1.00	0.00
More than one race, non-Hispanic	476	0.66	0.38	473	0.94	0.18	152	0.96	0.16	397	0.99	0.06	379	1.00	0.00
Socioeconomic status															
First quintile	2,594	0.45	0.38	2,917	0.86	0.26	753	0.91	0.22	2,363	0.99	0.08	1,964	1.00	0.00
Second quintile	3,271	0.61	0.37	3,503	0.91	0.22	925	0.95	0.16	2,796	0.99	0.05	2,230	1.00	0.00
Third quintile	3,470	0.68	0.36	3,686	0.94	0.17	997	0.98	0.12	3,003	1.00	0.04	2,437	1.00	0.00
Fourth quintile	3,650	0.77	0.32	3,909	0.97	0.11	1,019	0.99	0.07	3,173	1.00	0.03	2,688	1.00	0.00
Fifth quintile	3,880	0.86	0.26	4,152	0.98	0.09	1,159	1.00	0.04	3,642	1.00	0.01	3,158	1.00	0.00
School type															
Public school	13,736	0.65	0.37	14,578	0.93	0.20	3,809	0.96	0.14	12,998	0.99	0.05	11,575	1.00	0.00
Private school	3,888	0.84	0.28	4,357	0.97	0.12	1,042	0.99	0.06	3,279	1.00	0.03	2,623	1.00	0.00

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A18. Probability of proficiency, reading level 2: beginning sounds (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.29	0.33	18,935	0.68	0.33	5,053	0.81	0.28	16,336	0.96	0.14	14,246	1.00	0.00
Sex															
Male	8,984	0.27	0.32	9,688	0.64	0.35	2,556	0.78	0.30	8,349	0.95	0.15	7,204	1.00	0.00
Female	8,640	0.32	0.34	9,247	0.72	0.31	2,497	0.84	0.25	7,987	0.97	0.12	7,042	1.00	0.00
Race/ethnicity															
White, non-Hispanic	10,433	0.34	0.34	11,073	0.74	0.30	2,935	0.86	0.24	9,435	0.97	0.11	8,082	1.00	0.00
Black, non-Hispanic	2,854	0.21	0.28	2,968	0.57	0.35	782	0.74	0.30	2,371	0.93	0.18	1,840	1.00	0.00
Hispanic, race specified	1,182	0.23	0.31	1,315	0.63	0.35	322	0.78	0.28	1,233	0.96	0.13	1,252	1.00	0.00
Hispanic, race not specified	1,195	0.18	0.27	1,423	0.56	0.36	377	0.67	0.34	1,335	0.94	0.16	1,314	1.00	0.00
Asian	896	0.41	0.37	1,088	0.78	0.28	257	0.86	0.23	1,042	0.97	0.12	956	1.00	0.00
Hawaiian, other Pacific Islander	186	0.26	0.32	202	0.59	0.36	93	0.67	0.32	188	0.98	0.06	171	1.00	0.00
American Indian/Alaska Native	354	0.12	0.23	344	0.49	0.36	126	0.51	0.35	298	0.90	0.20	232	1.00	0.01
More than one race, non-Hispanic	476	0.28	0.33	473	0.66	0.33	152	0.82	0.27	397	0.96	0.15	379	1.00	0.00
Socioeconomic status															
First quintile	2,594	0.12	0.20	2,917	0.48	0.35	753	0.63	0.34	2,363	0.91	0.21	1,964	1.00	0.00
Second quintile	3,271	0.20	0.27	3,503	0.61	0.34	925	0.75	0.30	2,796	0.95	0.15	2,230	1.00	0.00
Third quintile	3,470	0.27	0.30	3,686	0.69	0.32	997	0.84	0.24	3,003	0.97	0.10	2,437	1.00	0.00
Fourth quintile	3,650	0.36	0.34	3,909	0.76	0.28	1,019	0.88	0.20	3,173	0.98	0.08	2,688	1.00	0.00
Fifth quintile	3,880	0.50	0.36	4,152	0.84	0.24	1,159	0.92	0.16	3,642	0.99	0.05	3,158	1.00	0.00
School type															
Public school	13,736	0.27	0.32	14,578	0.66	0.34	3,809	0.79	0.28	12,998	0.96	0.14	11,575	1.00	0.00
Private school	3,888	0.45	0.36	4,357	0.80	0.27	1,042	0.92	0.16	3,279	0.99	0.07	2,623	1.00	0.00

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A19. Probability of proficiency, reading level 3: ending sounds (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.16	0.26	18,935	0.49	0.34	5,053	0.65	0.32	16,336	0.91	0.20	14,246	1.00	0.01
Sex															
Male	8,984	0.15	0.25	9,688	0.46	0.35	2,556	0.61	0.34	8,349	0.90	0.21	7,204	1.00	0.01
Female	8,640	0.18	0.26	9,247	0.52	0.34	2,497	0.68	0.31	7,987	0.93	0.17	7,042	1.00	0.01
Race/ethnicity															
White, non-Hispanic	10,433	0.19	0.27	11,073	0.54	0.33	2,935	0.71	0.29	9,435	0.94	0.16	8,082	1.00	0.01
Black, non-Hispanic	2,854	0.10	0.20	2,968	0.38	0.34	782	0.55	0.34	2,371	0.85	0.25	1,840	1.00	0.02
Hispanic, race specified	1,182	0.12	0.23	1,315	0.44	0.34	322	0.61	0.34	1,233	0.90	0.20	1,252	1.00	0.01
Hispanic, race not specified	1,195	0.09	0.18	1,423	0.37	0.33	377	0.49	0.35	1,335	0.86	0.23	1,314	0.99	0.02
Asian	896	0.26	0.33	1,088	0.61	0.33	257	0.71	0.31	1,042	0.94	0.17	956	1.00	0.01
Hawaiian, other Pacific Islander	186	0.15	0.25	202	0.41	0.35	93	0.47	0.34	188	0.92	0.15	171	1.00	0.01
American Indian/Alaska Native	354	0.06	0.15	344	0.31	0.32	126	0.32	0.31	298	0.79	0.28	232	0.99	0.03
More than one race, non-Hispanic	476	0.16	0.26	473	0.46	0.34	152	0.65	0.31	397	0.92	0.19	379	1.00	0.01
Socioeconomic status															
First quintile	2,594	0.05	0.12	2,917	0.29	0.30	753	0.43	0.33	2,363	0.82	0.27	1,964	0.99	0.02
Second quintile	3,271	0.10	0.19	3,503	0.41	0.33	925	0.56	0.33	2,796	0.90	0.21	2,230	1.00	0.01
Third quintile	3,470	0.14	0.22	3,686	0.49	0.33	997	0.68	0.30	3,003	0.93	0.15	2,437	1.00	0.01
Fourth quintile	3,650	0.20	0.27	3,909	0.57	0.32	1,019	0.74	0.27	3,173	0.96	0.13	2,688	1.00	0.01
Fifth quintile	3,880	0.32	0.33	4,152	0.68	0.30	1,159	0.81	0.24	3,642	0.97	0.09	3,158	1.00	0.00
School type															
Public school	13,736	0.14	0.24	14,578	0.46	0.34	3,809	0.63	0.33	12,998	0.90	0.20	11,575	1.00	0.01
Private school	3,888	0.27	0.31	4,357	0.64	0.32	1,042	0.80	0.24	3,279	0.96	0.11	2,623	1.00	0.00

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A20. Probability of proficiency, reading level 4: sight words (range of possible values: 0.0 to 1.0) : School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.03	0.13	18,935	0.14	0.26	5,053	0.26	0.33	16,336	0.74	0.34	14,246	0.98	0.08
Sex															
Male	8,984	0.03	0.13	9,688	0.13	0.25	2,556	0.23	0.32	8,349	0.70	0.36	7,204	0.98	0.09
Female	8,640	0.03	0.12	9,247	0.16	0.27	2,497	0.29	0.35	7,987	0.78	0.32	7,042	0.99	0.07
Race/ethnicity															
White, non-Hispanic	10,433	0.03	0.14	11,073	0.16	0.27	2,935	0.30	0.34	9,435	0.80	0.30	8,082	0.99	0.05
Black, non-Hispanic	2,854	0.01	0.09	2,968	0.09	0.22	782	0.18	0.29	2,371	0.63	0.39	1,840	0.97	0.10
Hispanic, race specified	1,182	0.02	0.10	1,315	0.12	0.23	322	0.23	0.31	1,233	0.69	0.36	1,252	0.98	0.10
Hispanic, race not specified	1,195	0.01	0.07	1,423	0.08	0.20	377	0.14	0.26	1,335	0.61	0.38	1,314	0.97	0.11
Asian	896	0.08	0.24	1,088	0.26	0.36	257	0.41	0.43	1,042	0.81	0.31	956	0.99	0.05
Hawaiian, other Pacific Islander	186	0.03	0.13	202	0.12	0.24	93	0.14	0.28	188	0.69	0.35	171	0.99	0.09
American Indian/Alaska Native	354	0.01	0.05	344	0.06	0.16	126	0.06	0.17	298	0.48	0.39	232	0.95	0.16
More than one race, non-Hispanic	476	0.04	0.16	473	0.13	0.26	152	0.27	0.35	397	0.77	0.32	379	0.99	0.06
Socioeconomic status															
First quintile	2,594	0.00	0.05	2,917	0.05	0.14	753	0.10	0.21	2,363	0.54	0.39	1,964	0.96	0.14
Second quintile	3,271	0.01	0.08	3,503	0.09	0.21	925	0.18	0.28	2,796	0.70	0.36	2,230	0.98	0.09
Third quintile	3,470	0.02	0.10	3,686	0.12	0.23	997	0.26	0.33	3,003	0.77	0.32	2,437	0.99	0.05
Fourth quintile	3,650	0.03	0.13	3,909	0.17	0.28	1,019	0.31	0.34	3,173	0.83	0.27	2,688	1.00	0.04
Fifth quintile	3,880	0.07	0.20	4,152	0.27	0.34	1,159	0.43	0.38	3,642	0.88	0.23	3,158	1.00	0.01
School type															
Public school	13,736	0.02	0.11	14,578	0.12	0.24	3,809	0.24	0.32	12,998	0.72	0.35	11,575	0.98	0.08
Private school	3,888	0.06	0.18	4,357	0.24	0.33	1,042	0.40	0.37	3,279	0.86	0.25	2,623	1.00	0.02

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A21. Probability of proficiency, reading level 5: words in context (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.01	0.08	18,935	0.04	0.16	5,053	0.09	0.25	16,336	0.42	0.41	14,246	0.93	0.20
Sex															
Male	8,984	0.01	0.09	9,688	0.03	0.15	2,556	0.08	0.24	8,349	0.39	0.41	7,204	0.92	0.23
Female	8,640	0.01	0.08	9,247	0.04	0.16	2,497	0.10	0.26	7,987	0.46	0.41	7,042	0.95	0.17
Race/ethnicity															
White, non-Hispanic	10,433	0.01	0.09	11,073	0.04	0.18	2,935	0.11	0.27	9,435	0.49	0.41	8,082	0.96	0.15
Black, non-Hispanic	2,854	0.00	0.05	2,968	0.02	0.10	782	0.05	0.17	2,371	0.31	0.37	1,840	0.89	0.26
Hispanic, race specified	1,182	0.01	0.07	1,315	0.02	0.12	322	0.06	0.19	1,233	0.36	0.39	1,252	0.91	0.23
Hispanic, race not specified	1,195	0.00	0.04	1,423	0.01	0.09	377	0.03	0.14	1,335	0.27	0.36	1,314	0.86	0.28
Asian	896	0.04	0.18	1,088	0.10	0.26	257	0.24	0.38	1,042	0.56	0.42	956	0.97	0.13
Hawaiian, other Pacific Islander	186	0.01	0.07	202	0.01	0.08	93	0.05	0.18	188	0.34	0.39	171	0.94	0.16
American Indian/Alaska Native	354	0.00	0.01	344	0.01	0.07	126	0.01	0.09	298	0.18	0.31	232	0.81	0.32
More than one race, non-Hispanic	476	0.02	0.12	473	0.05	0.19	152	0.08	0.22	397	0.43	0.40	379	0.94	0.18
Socioeconomic status															
First quintile	2,594	0.00	0.03	2,917	0.00	0.05	753	0.02	0.10	2,363	0.21	0.32	1,964	0.84	0.30
Second quintile	3,271	0.00	0.05	3,503	0.02	0.12	925	0.04	0.17	2,796	0.36	0.39	2,230	0.92	0.22
Third quintile	3,470	0.00	0.05	3,686	0.02	0.12	997	0.08	0.24	3,003	0.43	0.40	2,437	0.96	0.16
Fourth quintile	3,650	0.01	0.08	3,909	0.04	0.17	1,019	0.11	0.27	3,173	0.50	0.40	2,688	0.97	0.12
Fifth quintile	3,880	0.03	0.14	4,152	0.09	0.25	1,159	0.18	0.34	3,642	0.63	0.39	3,158	0.99	0.07
School type															
Public school	13,736	0.01	0.07	14,578	0.03	0.14	3,809	0.08	0.23	12,998	0.40	0.40	11,575	0.93	0.21
Private school	3,888	0.02	0.12	4,357	0.07	0.22	1,042	0.15	0.31	3,279	0.60	0.40	2,623	0.98	0.08

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A22. Probability of proficiency, reading level 6: literal inference (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.00	0.04	18,935	0.01	0.07	5,053	0.03	0.14	16,336	0.14	0.28	14,246	0.74	0.35
Sex															
Male	8,984	0.00	0.04	9,688	0.01	0.07	2,556	0.03	0.13	8,349	0.13	0.27	7,204	0.72	0.37
Female	8,640	0.00	0.04	9,247	0.01	0.07	2,497	0.03	0.14	7,987	0.15	0.29	7,042	0.77	0.33
Race/ethnicity															
White, non-Hispanic	10,433	0.00	0.04	11,073	0.01	0.08	2,935	0.04	0.16	9,435	0.18	0.31	8,082	0.83	0.30
Black, non-Hispanic	2,854	0.00	0.02	2,968	0.00	0.03	782	0.01	0.08	2,371	0.07	0.19	1,840	0.61	0.39
Hispanic, race specified	1,182	0.00	0.05	1,315	0.01	0.06	322	0.01	0.09	1,233	0.10	0.24	1,252	0.67	0.38
Hispanic, race not specified	1,195	0.00	0.02	1,423	0.00	0.03	377	0.01	0.07	1,335	0.06	0.18	1,314	0.57	0.40
Asian	896	0.01	0.06	1,088	0.03	0.13	257	0.08	0.23	1,042	0.22	0.34	956	0.79	0.32
Hawaiian, other Pacific Islander	186	0.00	0.02	202	0.00	0.03	93	0.00	0.03	188	0.11	0.24	171	0.67	0.37
American Indian/Alaska Native	354	0.00	0.00	344	0.00	0.00	126	0.00	0.00	298	0.03	0.14	232	0.47	0.40
More than one race, non-Hispanic	476	0.00	0.04	473	0.01	0.10	152	0.02	0.11	397	0.14	0.29	379	0.76	0.35
Socioeconomic status															
First quintile	2,594	0.00	0.01	2,917	0.00	0.02	753	0.00	0.04	2,363	0.04	0.14	1,964	0.51	0.40
Second quintile	3,271	0.00	0.02	3,503	0.00	0.05	925	0.01	0.08	2,796	0.09	0.22	2,230	0.70	0.36
Third quintile	3,470	0.00	0.01	3,686	0.00	0.04	997	0.02	0.12	3,003	0.12	0.26	2,437	0.78	0.32
Fourth quintile	3,650	0.00	0.05	3,909	0.01	0.07	1,019	0.03	0.15	3,173	0.17	0.30	2,688	0.85	0.28
Fifth quintile	3,880	0.01	0.07	4,152	0.03	0.13	1,159	0.07	0.22	3,642	0.28	0.37	3,158	0.93	0.20
School type															
Public school	13,736	0.00	0.03	14,578	0.01	0.06	3,809	0.03	0.13	12,998	0.12	0.26	11,575	0.73	0.36
Private school	3,888	0.01	0.06	4,357	0.02	0.11	1,042	0.05	0.17	3,279	0.25	0.35	2,623	0.87	0.25

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A23. Probability of proficiency, reading level 7: extrapolation (range of possible values: 0.0 to 1.0) : School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.00	0.01	18,935	0.00	0.02	5,053	0.01	0.06	16,336	0.03	0.12	14,246	0.42	0.37
Sex															
Male	8,984	0.00	0.01	9,688	0.00	0.03	2,556	0.01	0.06	8,349	0.03	0.12	7,204	0.39	0.37
Female	8,640	0.00	0.01	9,247	0.00	0.02	2,497	0.01	0.05	7,987	0.04	0.12	7,042	0.45	0.38
Race/ethnicity															
White, non-Hispanic	10,433	0.00	0.01	11,073	0.00	0.03	2,935	0.01	0.06	9,435	0.05	0.14	8,082	0.52	0.37
Black, non-Hispanic	2,854	0.00	0.01	2,968	0.00	0.00	782	0.00	0.03	2,371	0.01	0.06	1,840	0.26	0.31
Hispanic, race specified	1,182	0.00	0.01	1,315	0.00	0.02	322	0.01	0.06	1,233	0.02	0.08	1,252	0.34	0.36
Hispanic, race not specified	1,195	0.00	0.00	1,423	0.00	0.01	377	0.00	0.02	1,335	0.01	0.06	1,314	0.24	0.31
Asian	896	0.00	0.01	1,088	0.00	0.03	257	0.02	0.10	1,042	0.06	0.16	956	0.46	0.37
Hawaiian, other Pacific Islander	186	0.00	0.00	202	0.00	0.00	93	0.00	0.00	188	0.02	0.10	171	0.32	0.34
American Indian/Alaska Native	354	0.00	0.00	344	0.00	0.00	126	0.00	0.00	298	0.01	0.07	232	0.18	0.29
More than one race, non-Hispanic	476	0.00	0.01	473	0.00	0.03	152	0.00	0.02	397	0.04	0.15	379	0.43	0.36
Socioeconomic status															
First quintile	2,594	0.00	0.00	2,917	0.00	0.01	753	0.00	0.02	2,363	0.01	0.04	1,964	0.19	0.27
Second quintile	3,271	0.00	0.00	3,503	0.00	0.01	925	0.00	0.04	2,796	0.02	0.08	2,230	0.34	0.34
Third quintile	3,470	0.00	0.00	3,686	0.00	0.01	997	0.00	0.03	3,003	0.03	0.09	2,437	0.43	0.36
Fourth quintile	3,650	0.00	0.01	3,909	0.00	0.02	1,019	0.01	0.05	3,173	0.04	0.12	2,688	0.53	0.37
Fifth quintile	3,880	0.00	0.02	4,152	0.01	0.05	1,159	0.02	0.10	3,642	0.09	0.20	3,158	0.68	0.33
School type															
Public school	13,736	0.00	0.01	14,578	0.00	0.02	3,809	0.01	0.05	12,998	0.03	0.11	11,575	0.40	0.37
Private school	3,888	0.00	0.01	4,357	0.00	0.04	1,042	0.01	0.07	3,279	0.06	0.16	2,623	0.57	0.36

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A24. Probability of proficiency, reading level 8: evaluation (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.00	0.00	18,935	0.00	0.01	5,053	0.00	0.03	16,336	0.02	0.07	14,246	0.26	0.27
Sex															
Male	8,984	0.00	0.00	9,688	0.00	0.01	2,556	0.00	0.03	8,349	0.02	0.07	7,204	0.24	0.26
Female	8,640	0.00	0.00	9,247	0.00	0.01	2,497	0.00	0.03	7,987	0.02	0.07	7,042	0.28	0.27
Race/ethnicity															
White, non-Hispanic	10,433	0.00	0.00	11,073	0.00	0.02	2,935	0.01	0.04	9,435	0.03	0.08	8,082	0.33	0.28
Black, non-Hispanic	2,854	0.00	0.00	2,968	0.00	0.00	782	0.00	0.01	2,371	0.01	0.04	1,840	0.15	0.19
Hispanic, race specified	1,182	0.00	0.01	1,315	0.00	0.01	322	0.00	0.03	1,233	0.01	0.05	1,252	0.20	0.24
Hispanic, race not specified	1,195	0.00	0.00	1,423	0.00	0.01	377	0.00	0.01	1,335	0.01	0.03	1,314	0.14	0.19
Asian	896	0.00	0.01	1,088	0.00	0.02	257	0.01	0.05	1,042	0.04	0.09	956	0.28	0.27
Hawaiian, other Pacific Islander	186	0.00	0.00	202	0.00	0.00	93	0.00	0.00	188	0.02	0.06	171	0.19	0.22
American Indian/Alaska Native	354	0.00	0.00	344	0.00	0.00	126	0.00	0.00	298	0.01	0.04	232	0.11	0.18
More than one race, non-Hispanic	476	0.00	0.01	473	0.00	0.01	152	0.00	0.01	397	0.03	0.09	379	0.26	0.26
Socioeconomic status															
First quintile	2,594	0.00	0.00	2,917	0.00	0.00	753	0.00	0.01	2,363	0.01	0.02	1,964	0.11	0.16
Second quintile	3,271	0.00	0.00	3,503	0.00	0.01	925	0.00	0.02	2,796	0.01	0.04	2,230	0.20	0.22
Third quintile	3,470	0.00	0.00	3,686	0.00	0.01	997	0.00	0.02	3,003	0.02	0.05	2,437	0.26	0.25
Fourth quintile	3,650	0.00	0.01	3,909	0.00	0.01	1,019	0.00	0.03	3,173	0.03	0.07	2,688	0.33	0.28
Fifth quintile	3,880	0.00	0.01	4,152	0.00	0.03	1,159	0.01	0.05	3,642	0.05	0.12	3,158	0.45	0.29
School type															
Public school	13,736	0.00	0.00	14,578	0.00	0.01	3,809	0.00	0.03	12,998	0.02	0.06	11,575	0.25	0.26
Private school	3,888	0.00	0.01	4,357	0.00	0.02	1,042	0.01	0.04	3,279	0.04	0.09	2,623	0.37	0.29

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A25. Probability of proficiency, mathematics level 1: count, number, shape (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.91	0.19	19,647	0.99	0.07	5,226	0.99	0.05	16,641	1.00	0.02	14,349	1.00	0.00
Sex															
Male	9,479	0.91	0.20	10,041	0.98	0.07	2,644	0.99	0.05	8,506	1.00	0.03	7,277	1.00	0.00
Female	9,156	0.92	0.18	9,606	0.99	0.07	2,582	0.99	0.04	8,135	1.00	0.02	7,072	1.00	0.00
Race/ethnicity															
White, non-Hispanic	10,433	0.95	0.13	11,071	0.99	0.05	2,935	1.00	0.03	9,436	1.00	0.02	8,116	1.00	0.00
Black, non-Hispanic	2,855	0.88	0.22	2,962	0.98	0.09	781	0.99	0.07	2,371	1.00	0.03	1,871	1.00	0.00
Hispanic, race specified	1,588	0.86	0.23	1,624	0.98	0.08	389	0.99	0.04	1,354	1.00	0.00	1,260	1.00	0.00
Hispanic, race not specified	1,800	0.81	0.27	1,834	0.97	0.11	486	0.98	0.06	1,518	1.00	0.02	1,324	1.00	0.00
Asian	897	0.96	0.12	1,088	1.00	0.03	256	1.00	0.01	1,042	1.00	0.00	956	1.00	0.00
Hawaiian, other Pacific Islander	187	0.90	0.22	202	0.98	0.06	93	1.00	0.00	188	1.00	0.00	172	1.00	0.00
American Indian/Alaska Native	354	0.81	0.27	345	0.97	0.08	126	0.97	0.12	298	1.00	0.00	250	1.00	0.00
More than one race, non-Hispanic	473	0.93	0.15	472	0.99	0.06	151	0.99	0.04	397	1.00	0.03	380	1.00	0.00
Socioeconomic status															
First quintile	3,269	0.80	0.27	3,426	0.96	0.11	895	0.98	0.06	2,572	1.00	0.03	2,001	1.00	0.00
Second quintile	3,429	0.89	0.21	3,607	0.98	0.08	942	0.99	0.07	2,839	1.00	0.04	2,250	1.00	0.00
Third quintile	3,546	0.94	0.15	3,721	0.99	0.05	1,001	0.99	0.04	3,017	1.00	0.00	2,452	1.00	0.00
Fourth quintile	3,676	0.96	0.12	3,921	0.99	0.04	1,023	1.00	0.02	3,178	1.00	0.01	2,693	1.00	0.00
Fifth quintile	3,893	0.98	0.08	4,161	1.00	0.04	1,158	1.00	0.00	3,644	1.00	0.01	3,163	1.00	0.00
School type															
Public school	14,701	0.90	0.20	15,259	0.98	0.07	3,971	0.99	0.05	13,292	1.00	0.02	11,670	1.00	0.00
Private school	3,934	0.97	0.11	4,388	0.99	0.05	1,043	1.00	0.00	3,286	1.00	0.02	2,631	1.00	0.00

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A26. Probability of proficiency, mathematics level 2: relative size (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.53	0.35	19,647	0.83	0.25	5,226	0.91	0.19	16,641	0.98	0.09	14,349	1.00	0.00
Sex															
Male	9,479	0.53	0.36	10,041	0.82	0.26	2,644	0.90	0.21	8,506	0.98	0.10	7,277	1.00	0.00
Female	9,156	0.54	0.34	9,606	0.83	0.24	2,582	0.92	0.17	8,135	0.98	0.08	7,072	1.00	0.00
Race/ethnicity															
White, non-Hispanic	10,433	0.63	0.33	11,071	0.89	0.20	2,935	0.95	0.14	9,436	0.99	0.08	8,116	1.00	0.00
Black, non-Hispanic	2,855	0.41	0.33	2,962	0.74	0.28	781	0.86	0.23	2,371	0.96	0.12	1,871	1.00	0.00
Hispanic, race specified	1,588	0.41	0.34	1,624	0.76	0.29	389	0.88	0.21	1,354	0.98	0.09	1,260	1.00	0.00
Hispanic, race not specified	1,800	0.32	0.32	1,834	0.69	0.31	486	0.82	0.26	1,518	0.98	0.09	1,324	1.00	0.00
Asian	897	0.66	0.32	1,088	0.89	0.18	256	0.94	0.12	1,042	0.99	0.06	956	1.00	0.00
Hawaiian, other Pacific Islander	187	0.47	0.33	202	0.78	0.28	93	0.89	0.16	188	0.98	0.05	172	1.00	0.00
American Indian/Alaska Native	354	0.34	0.33	345	0.73	0.29	126	0.76	0.30	298	0.97	0.11	250	1.00	0.00
More than one race, non-Hispanic	473	0.53	0.34	472	0.84	0.23	151	0.90	0.21	397	0.98	0.10	380	1.00	0.00
Socioeconomic status															
First quintile	3,269	0.30	0.30	3,426	0.68	0.31	895	0.79	0.27	2,572	0.96	0.13	2,001	1.00	0.00
Second quintile	3,429	0.45	0.34	3,607	0.80	0.26	942	0.89	0.20	2,839	0.98	0.11	2,250	1.00	0.00
Third quintile	3,546	0.55	0.33	3,721	0.86	0.21	1,001	0.94	0.15	3,017	0.99	0.08	2,452	1.00	0.00
Fourth quintile	3,676	0.64	0.32	3,921	0.90	0.18	1,023	0.96	0.12	3,178	0.99	0.05	2,693	1.00	0.00
Fifth quintile	3,893	0.75	0.28	4,161	0.94	0.14	1,158	0.97	0.08	3,644	1.00	0.03	3,163	1.00	0.00
School type															
Public school	14,701	0.51	0.35	15,259	0.81	0.26	3,971	0.90	0.20	13,292	0.98	0.09	11,670	1.00	0.00
Private school	3,934	0.70	0.30	4,388	0.91	0.17	1,043	0.98	0.07	3,286	0.99	0.05	2,631	1.00	0.00

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A27. Probability of proficiency, mathematics level 3: ordinality, sequence (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.19	0.30	19,647	0.53	0.39	5,226	0.71	0.36	16,641	0.93	0.20	14,349	1.00	0.01
Sex															
Male	9,479	0.20	0.31	10,041	0.52	0.39	2,644	0.70	0.37	8,506	0.93	0.21	7,277	1.00	0.01
Female	9,156	0.19	0.29	9,606	0.53	0.39	2,582	0.72	0.35	8,135	0.93	0.19	7,072	1.00	0.01
Race/ethnicity															
White, non-Hispanic	10,433	0.26	0.33	11,071	0.63	0.37	2,935	0.80	0.30	9,436	0.96	0.16	8,116	1.00	0.01
Black, non-Hispanic	2,855	0.09	0.20	2,962	0.37	0.37	781	0.59	0.39	2,371	0.87	0.27	1,871	1.00	0.01
Hispanic, race specified	1,588	0.11	0.23	1,624	0.41	0.39	389	0.65	0.38	1,354	0.90	0.24	1,260	1.00	0.01
Hispanic, race not specified	1,800	0.07	0.18	1,834	0.33	0.37	486	0.52	0.40	1,518	0.90	0.23	1,324	1.00	0.02
Asian	897	0.30	0.36	1,088	0.62	0.38	256	0.76	0.35	1,042	0.94	0.17	956	1.00	0.00
Hawaiian, other Pacific Islander	187	0.12	0.24	202	0.41	0.37	93	0.57	0.36	188	0.91	0.21	172	1.00	0.02
American Indian/Alaska Native	354	0.08	0.19	345	0.35	0.37	126	0.43	0.40	298	0.89	0.24	250	1.00	0.01
More than one race, non-Hispanic	473	0.18	0.29	472	0.51	0.38	151	0.72	0.35	397	0.94	0.19	380	1.00	0.00
Socioeconomic status															
First quintile	3,269	0.05	0.15	3,426	0.28	0.34	895	0.46	0.40	2,572	0.86	0.28	2,001	1.00	0.02
Second quintile	3,429	0.12	0.23	3,607	0.46	0.38	942	0.66	0.37	2,839	0.91	0.22	2,250	1.00	0.01
Third quintile	3,546	0.17	0.27	3,721	0.55	0.38	1,001	0.77	0.31	3,017	0.95	0.17	2,452	1.00	0.01
Fourth quintile	3,676	0.25	0.32	3,921	0.63	0.36	1,023	0.82	0.28	3,178	0.97	0.13	2,693	1.00	0.00
Fifth quintile	3,893	0.38	0.37	4,161	0.75	0.32	1,158	0.88	0.24	3,644	0.98	0.09	3,163	1.00	0.00
School type															
Public school	14,701	0.17	0.28	15,259	0.50	0.39	3,971	0.69	0.37	13,292	0.92	0.21	11,670	1.00	0.01
Private school	3,934	0.32	0.36	4,388	0.68	0.35	1,043	0.88	0.22	3,286	0.98	0.10	2,631	1.00	0.00

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A28. Probability of proficiency, mathematics level 4: add/subtract (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.03	0.11	19,647	0.16	0.25	5,226	0.31	0.33	16,641	0.69	0.33	14,349	0.96	0.11
Sex															
Male	9,479	0.04	0.13	10,041	0.16	0.26	2,644	0.32	0.34	8,506	0.69	0.34	7,277	0.96	0.11
Female	9,156	0.03	0.10	9,606	0.15	0.24	2,582	0.30	0.32	8,135	0.69	0.32	7,072	0.96	0.10
Race/ethnicity															
White, non-Hispanic	10,433	0.05	0.14	11,071	0.21	0.28	2,935	0.38	0.34	9,436	0.77	0.29	8,116	0.98	0.08
Black, non-Hispanic	2,855	0.01	0.05	2,962	0.07	0.16	781	0.20	0.27	2,371	0.55	0.35	1,871	0.93	0.14
Hispanic, race specified	1,588	0.02	0.07	1,624	0.10	0.19	389	0.25	0.30	1,354	0.63	0.35	1,260	0.95	0.12
Hispanic, race not specified	1,800	0.01	0.04	1,834	0.07	0.16	486	0.16	0.24	1,518	0.56	0.34	1,324	0.94	0.13
Asian	897	0.08	0.19	1,088	0.23	0.30	256	0.40	0.36	1,042	0.72	0.33	956	0.97	0.08
Hawaiian, other Pacific Islander	187	0.02	0.10	202	0.09	0.18	93	0.16	0.24	188	0.56	0.34	172	0.95	0.13
American Indian/Alaska Native	354	0.01	0.05	345	0.07	0.16	126	0.12	0.21	298	0.52	0.35	250	0.94	0.12
More than one race, non-Hispanic	473	0.03	0.11	472	0.14	0.23	151	0.26	0.29	397	0.69	0.34	380	0.96	0.11
Socioeconomic status															
First quintile	3,269	0.01	0.04	3,426	0.05	0.13	895	0.14	0.23	2,572	0.51	0.35	2,001	0.92	0.16
Second quintile	3,429	0.01	0.06	3,607	0.11	0.19	942	0.22	0.28	2,839	0.64	0.34	2,250	0.96	0.11
Third quintile	3,546	0.02	0.08	3,721	0.14	0.22	1,001	0.31	0.31	3,017	0.72	0.31	2,452	0.97	0.08
Fourth quintile	3,676	0.04	0.12	3,921	0.20	0.26	1,023	0.37	0.32	3,178	0.78	0.28	2,693	0.98	0.07
Fifth quintile	3,893	0.09	0.19	4,161	0.30	0.32	1,158	0.52	0.36	3,644	0.86	0.23	3,163	0.99	0.03
School type															
Public school	14,701	0.03	0.10	15,259	0.14	0.23	3,971	0.29	0.32	13,292	0.68	0.34	11,670	0.96	0.11
Private school	3,934	0.07	0.17	4,388	0.25	0.30	1,043	0.45	0.33	3,286	0.81	0.26	2,631	0.98	0.05

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A29. Probability of proficiency, mathematics level 5: multiply/divide (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.00	0.03	19,647	0.01	0.08	5,226	0.04	0.14	16,641	0.22	0.30	14,349	0.75	0.32
Sex															
Male	9,479	0.00	0.04	10,041	0.02	0.09	2,644	0.05	0.16	8,506	0.24	0.31	7,277	0.77	0.32
Female	9,156	0.00	0.01	9,606	0.01	0.05	2,582	0.04	0.12	8,135	0.20	0.27	7,072	0.74	0.32
Race/ethnicity															
White, non-Hispanic	10,433	0.00	0.04	11,071	0.02	0.09	2,935	0.06	0.17	9,436	0.29	0.32	8,116	0.83	0.27
Black, non-Hispanic	2,855	0.00	0.02	2,962	0.00	0.03	781	0.01	0.07	2,371	0.09	0.18	1,871	0.58	0.37
Hispanic, race specified	1,588	0.00	0.02	1,624	0.01	0.04	389	0.02	0.08	1,354	0.16	0.25	1,260	0.69	0.34
Hispanic, race not specified	1,800	0.00	0.00	1,834	0.00	0.03	486	0.01	0.03	1,518	0.10	0.19	1,324	0.65	0.35
Asian	897	0.01	0.05	1,088	0.03	0.13	256	0.09	0.22	1,042	0.27	0.32	956	0.80	0.31
Hawaiian, other Pacific Islander	187	0.00	0.00	202	0.01	0.05	93	0.01	0.04	188	0.10	0.19	172	0.71	0.33
American Indian/Alaska Native	354	0.00	0.00	345	0.00	0.04	126	0.01	0.05	298	0.09	0.18	250	0.57	0.36
More than one race, non-Hispanic	473	0.00	0.04	472	0.01	0.07	151	0.03	0.11	397	0.23	0.30	380	0.78	0.31
Socioeconomic status															
First quintile	3,269	0.00	0.01	3,426	0.00	0.02	895	0.01	0.04	2,572	0.08	0.17	2,001	0.55	0.37
Second quintile	3,429	0.00	0.00	3,607	0.01	0.04	942	0.02	0.10	2,839	0.15	0.24	2,250	0.71	0.33
Third quintile	3,546	0.00	0.01	3,721	0.01	0.06	1,001	0.03	0.10	3,017	0.21	0.28	2,452	0.79	0.29
Fourth quintile	3,676	0.00	0.03	3,921	0.02	0.08	1,023	0.05	0.13	3,178	0.28	0.31	2,693	0.85	0.25
Fifth quintile	3,893	0.01	0.06	4,161	0.04	0.13	1,158	0.12	0.23	3,644	0.41	0.35	3,163	0.92	0.19
School type															
Public school	14,701	0.00	0.03	15,259	0.01	0.07	3,971	0.04	0.13	13,292	0.21	0.29	11,670	0.75	0.33
Private school	3,934	0.01	0.05	4,388	0.03	0.11	1,043	0.08	0.19	3,286	0.33	0.33	2,631	0.84	0.26

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A30. Probability of proficiency, mathematics level 6: place value (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.00	0.01	19,647	0.00	0.01	5,226	0.00	0.03	16,641	0.03	0.11	14,349	0.39	0.39
Sex															
Male	9,479	0.00	0.01	10,041	0.00	0.02	2,644	0.00	0.04	8,506	0.03	0.12	7,277	0.43	0.40
Female	9,156	0.00	0.00	9,606	0.00	0.00	2,582	0.00	0.01	8,135	0.02	0.08	7,072	0.35	0.38
Race/ethnicity															
White, non-Hispanic	10,433	0.00	0.00	11,071	0.00	0.02	2,935	0.00	0.03	9,436	0.04	0.13	8,116	0.49	0.39
Black, non-Hispanic	2,855	0.00	0.00	2,962	0.00	0.01	781	0.00	0.01	2,371	0.01	0.04	1,871	0.19	0.30
Hispanic, race specified	1,588	0.00	0.00	1,624	0.00	0.00	389	0.00	0.01	1,354	0.02	0.07	1,260	0.31	0.37
Hispanic, race not specified	1,800	0.00	0.00	1,834	0.00	0.00	486	0.00	0.00	1,518	0.00	0.03	1,324	0.24	0.33
Asian	897	0.00	0.00	1,088	0.00	0.03	256	0.01	0.05	1,042	0.05	0.16	956	0.50	0.41
Hawaiian, other Pacific Islander	187	0.00	0.00	202	0.00	0.00	93	0.00	0.00	188	0.00	0.03	172	0.26	0.32
American Indian/Alaska Native	354	0.00	0.00	345	0.00	0.00	126	0.00	0.00	298	0.01	0.05	250	0.19	0.31
More than one race, non-Hispanic	473	0.00	0.03	472	0.00	0.03	151	0.00	0.01	397	0.02	0.08	380	0.41	0.39
Socioeconomic status															
First quintile	3,269	0.00	0.00	3,426	0.00	0.00	895	0.00	0.00	2,572	0.01	0.04	2,001	0.17	0.28
Second quintile	3,429	0.00	0.00	3,607	0.00	0.00	942	0.00	0.02	2,839	0.01	0.07	2,250	0.30	0.35
Third quintile	3,546	0.00	0.00	3,721	0.00	0.01	1,001	0.00	0.01	3,017	0.02	0.09	2,452	0.39	0.37
Fourth quintile	3,676	0.00	0.00	3,921	0.00	0.01	1,023	0.00	0.02	3,178	0.03	0.10	2,693	0.51	0.39
Fifth quintile	3,893	0.00	0.01	4,161	0.00	0.03	1,158	0.01	0.06	3,644	0.07	0.17	3,163	0.65	0.37
School type															
Public school	14,701	0.00	0.01	15,259	0.00	0.01	3,971	0.00	0.03	13,292	0.02	0.10	11,670	0.38	0.39
Private school	3,934	0.00	0.00	4,388	0.00	0.02	1,043	0.01	0.04	3,286	0.04	0.13	2,631	0.48	0.39

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A31. Probability of proficiency, mathematics level 7: rate and measurement (range of possible values: 0.0 to 1.0) : School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.00	0.00	19,647	0.00	0.00	5,226	0.00	0.00	16,641	0.00	0.02	14,349	0.14	0.28
Sex															
Male	9,479	0.00	0.00	10,041	0.00	0.00	2,644	0.00	0.00	8,506	0.00	0.03	7,277	0.17	0.30
Female	9,156	0.00	0.00	9,606	0.00	0.00	2,582	0.00	0.00	8,135	0.00	0.02	7,072	0.11	0.24
Race/ethnicity															
White, non-Hispanic	10,433	0.00	0.00	11,071	0.00	0.00	2,935	0.00	0.00	9,436	0.00	0.03	8,116	0.19	0.31
Black, non-Hispanic	2,855	0.00	0.00	2,962	0.00	0.00	781	0.00	0.00	2,371	0.00	0.00	1,871	0.05	0.16
Hispanic, race specified	1,588	0.00	0.00	1,624	0.00	0.00	389	0.00	0.00	1,354	0.00	0.01	1,260	0.10	0.23
Hispanic, race not specified	1,800	0.00	0.00	1,834	0.00	0.00	486	0.00	0.00	1,518	0.00	0.00	1,324	0.06	0.18
Asian	897	0.00	0.00	1,088	0.00	0.00	256	0.00	0.00	1,042	0.00	0.02	956	0.23	0.33
Hawaiian, other Pacific Islander	187	0.00	0.00	202	0.00	0.00	93	0.00	0.00	188	0.00	0.00	172	0.06	0.18
American Indian/Alaska Native	354	0.00	0.00	345	0.00	0.00	126	0.00	0.00	298	0.00	0.02	250	0.04	0.15
More than one race, non-Hispanic	473	0.00	0.00	472	0.00	0.00	151	0.00	0.00	397	0.00	0.00	380	0.15	0.28
Socioeconomic status															
First quintile	3,269	0.00	0.00	3,426	0.00	0.00	895	0.00	0.00	2,572	0.00	0.00	2,001	0.04	0.13
Second quintile	3,429	0.00	0.00	3,607	0.00	0.00	942	0.00	0.00	2,839	0.00	0.01	2,250	0.08	0.20
Third quintile	3,546	0.00	0.00	3,721	0.00	0.00	1,001	0.00	0.00	3,017	0.00	0.01	2,452	0.12	0.24
Fourth quintile	3,676	0.00	0.00	3,921	0.00	0.00	1,023	0.00	0.00	3,178	0.00	0.02	2,693	0.20	0.31
Fifth quintile	3,893	0.00	0.00	4,161	0.00	0.00	1,158	0.00	0.00	3,644	0.01	0.04	3,163	0.32	0.37
School type															
Public school	14,701	0.00	0.00	15,259	0.00	0.00	3,971	0.00	0.00	13,292	0.00	0.02	11,670	0.14	0.27
Private school	3,934	0.00	0.00	4,388	0.00	0.00	1,043	0.00	0.00	3,286	0.00	0.03	2,631	0.19	0.31

[1] Number in sample.

[2] Standard deviation.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A32. Percent of children at or above modal reading proficiency for each grade: School years 1998–99, 1999–2000, and 2001–02

Characteristic	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	
	N[1]	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Total sample	16,739	64.5	17,691	48.6	4,740	65.2	15,226	77.6	13,259	71.5
Sex										
Male	8,536	60.7	9,003	45.2	2,394	60.5	7,736	73.4	6,738	68.2
Female	8,203	68.5	8,688	52.1	2,346	70.1	7,490	82.1	6,521	75.0
Race/ethnicity										
White, non-Hispanic	9,887	69.9	10,402	55.0	2,787	72.5	8,931	83.0	7,510	80.0
Black, non-Hispanic	2,744	59.9	2,735	33.9	714	51.7	2,129	67.6	1,718	60.2
Hispanic, race specified	1,126	51.8	1,225	44.3	295	63.0	1,142	73.9	1,160	62.7
Hispanic, race not specified	1,130	46.7	1,319	38.0	343	50.3	1,200	65.0	1,224	53.1
Asian	845	81.5	1,017	63.4	240	68.9	970	84.9	905	74.1
Hawaiian, other Pacific Islander	179	64	185	35.2	90	42.0	167	75.8	154	56.5
American Indian/Alaska Native	336	34.9	324	26.8	121	24.8	276	50.0	219	44.9
More than one race, non-Hispanic	447	62.7	438	44.6	142	65.3	378	82.3	349	74.4
Socioeconomic status										
First quintile	2,514	41.7	2,726	27.0	688	40.1	2,110	59.0	1,844	48.4
Second quintile	3,114	56.6	3,259	39.1	861	55.7	2,577	73.8	2,065	65.8
Third quintile	3,294	64.4	3,416	48.3	935	67.7	2,798	80.2	2,253	76.1
Fourth quintile	3,462	73.7	3,643	58.3	964	76.5	3,016	85.8	2,479	81.4
Fifth quintile	3,629	83.8	3,924	70.1	1,100	83.4	3,511	90.6	2,969	90.8
School type										
Public school	13,073	61.2	13,614	45.3	3,565	63.1	12,054	75.9	10,749	69.9
Private school	3,666	82.8	4,077	66.2	986	83.4	3,125	89.2	2,463	84.2

[1] Number in sample.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

Table A33. Percent of children at or above modal mathematics proficiency for each grade: School years 1998–99, 1999–2000, and 2001–02

Characteristic	Round 1 Modal Level =2		Round 2 Modal Level =3		Round 3 Modal Level =3		Round 4 Modal Level =4		Round 5 Modal Level =5	
	N[1]	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Total sample	18,149	55.4	18,945	53.4	5,060	70.7	16,133	70.2	13,998	74.3
Sex										
Male	9,178	52.9	9,641	52.6	2,550	69.8	8,199	70.2	7,081	75.7
Female	8,971	58.1	9,304	54.2	2,510	71.6	7,934	70.2	6,917	72.8
Race/ethnicity										
White, non-Hispanic	10,104	66.1	10,703	64.9	2,848	80.4	9,223	78.0	7,934	82.5
Black, non-Hispanic	2,800	44.5	2,838	37.2	753	60.0	2,287	58.0	1,837	53.6
Hispanic, race specified	1,552	40.4	1,555	38.6	374	61.5	1,293	62.2	1,223	68.1
Hispanic, race not specified	1,777	27.8	1,766	30.1	464	46.4	1,426	55.7	1,283	66.4
Asian	873	69.4	1,049	61.4	252	72.3	1,016	70.4	922	77.3
Hawaiian, other Pacific Islander	182	53.4	196	39.2	91	56.4	179	54.1	167	70.7
American Indian/Alaska Native	351	35.0	331	34.2	123	41.2	287	48.2	239	56.9
More than one race, non-Hispanic	463	57.8	459	51.9	146	70.7	385	69.6	373	80.5
Socioeconomic status										
First quintile	3,212	28.7	3,278	27.0	864	42.4	2,434	50.1	1,944	54.6
Second quintile	3,353	47.4	3,485	46.3	911	64.5	2,737	64.4	2,188	69.5
Third quintile	3,458	58.2	3,593	55.9	959	79.2	2,942	72.7	2,405	78.0
Fourth quintile	3,568	67.3	3,778	65.2	995	81.5	3,116	79.6	2,622	83.6
Fifth quintile	3,751	78.8	4,031	76.7	1,130	88.3	3,566	86.7	3,105	90.4
School type										
Public school	14,332	52.3	14,692	50.4	3,831	68.2	12,856	68.7	11,390	73.4
Private school	3,817	73.5	4,253	70.0	1,023	89.8	3,215	81.8	2,562	82.5

[1] Number in sample.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. 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 2002.

This page is intentionally left blank.

APPENDIX B

ECLS-K ITEM PARAMETERS AND ITEM FIT BY ROUNDS

Table B1. Reading assessment item parameters and item fit by rounds: School years 1998–99, 1999–2000, and 2001–02

Reading	Test Form(s) R1-R4	Test Form(s) R5	IRT parameters			Round 1				Round 2				Round 3				Round 4				Round 5			
							P+		Differ-		P+		Differ-		P+		Differ-		P+		Differ-		P+		Differ-
			a[1]	b[2]	c[3]	N[4]	Actual	Predicted	ence	N	Actual	Predicted	ence	N	Actual	Predicted	ence	N	Actual	Predicted	ence	N	Actual	Predicted	ence
LETRECD	R	†	2.43	-1.52	0.00	17607	0.74	0.72	0.02	18938	0.94	0.93	0.00	5053	0.96	0.96	0.00	16339	0.99	0.99	-0.01	†	†	†	†
LETRECF	R	†	2.76	-1.47	0.00	17606	0.72	0.70	0.01	18939	0.94	0.93	0.00	5053	0.96	0.97	0.00	16339	0.99	1.00	0.00	†	†	†	†
LETRECT	R	†	2.59	-1.38	0.00	17594	0.66	0.65	0.01	18937	0.92	0.91	0.01	5053	0.96	0.95	0.00	16339	0.99	0.99	0.00	†	†	†	†
LETRECM	R	†	2.43	-1.46	0.00	17604	0.71	0.69	0.01	18935	0.93	0.92	0.00	5053	0.95	0.96	-0.01	16338	0.99	0.99	0.00	†	†	†	†
BEGP	R	†	1.57	-1.01	0.00	17607	0.47	0.46	0.01	18939	0.77	0.75	0.02	5053	0.84	0.84	0.00	16340	0.93	0.96	-0.02	†	†	†	†
BEGR	R	†	2.10	-0.99	0.00	17600	0.42	0.43	-0.01	18936	0.79	0.77	0.02	5053	0.88	0.86	0.02	16340	0.96	0.97	-0.01	†	†	†	†
BEGL	R	†	2.07	-0.95	0.00	17605	0.41	0.41	-0.01	18940	0.78	0.75	0.02	5053	0.87	0.85	0.02	16340	0.96	0.97	-0.01	†	†	†	†
BEGB	R	†	1.29	-0.61	0.00	17609	0.32	0.31	0.01	18938	0.58	0.57	0.01	5052	0.68	0.68	0.00	16340	0.86	0.88	-0.02	†	†	†	†
ENDD	R	†	1.52	-0.40	0.00	17602	0.22	0.21	0.01	18937	0.48	0.48	0.00	5052	0.61	0.60	0.01	16338	0.86	0.86	0.00	†	†	†	†
ENDP	R	†	1.47	-0.55	0.00	17614	0.28	0.27	0.01	18939	0.55	0.55	0.00	5052	0.67	0.66	0.01	16338	0.88	0.88	-0.01	†	†	†	†
ENDL	R	†	1.95	-0.75	0.00	17599	0.32	0.32	0.00	18930	0.66	0.66	0.01	5051	0.78	0.77	0.00	16338	0.94	0.95	-0.01	†	†	†	†
ENDF	R	†	1.63	-0.70	0.00	17607	0.32	0.32	0.00	18938	0.64	0.63	0.01	5050	0.75	0.74	0.01	16338	0.91	0.92	-0.01	†	†	†	†
CEREAL	L	†	1.03	-2.68	0.00	13350	0.91	0.90	0.01	6516	0.94	0.94	0.00	1062	0.95	0.95	0.00	617	0.91	0.96	-0.04	†	†	†	†
BEGBIKE	L	†	1.48	-1.79	0.00	13347	0.73	0.73	0.01	6520	0.84	0.84	0.00	1062	0.87	0.86	0.01	618	0.88	0.89	-0.01	†	†	†	†
BEGIN	L , M , H	†	0.81	-1.70	0.00	17611	0.69	0.68	0.01	18936	0.83	0.83	0.00	5051	0.87	0.87	0.01	16338	0.93	0.94	-0.01	†	†	†	†
NEXTLINE	L , M , H	†	1.00	-1.22	0.00	17610	0.54	0.54	0.00	18939	0.76	0.75	0.00	5052	0.85	0.82	0.04	16339	0.92	0.93	-0.01	†	†	†	†
STORYEND	L , M , H	†	1.16	-1.21	0.00	17615	0.57	0.54	0.02	18940	0.78	0.78	0.00	5049	0.84	0.84	0.00	16336	0.92	0.95	-0.02	†	†	†	†
CANDLE	L	†	0.71	-3.54	0.17	13348	0.95	0.94	0.00	6519	0.95	0.96	-0.01	1061	0.96	0.96	0.00	617	0.94	0.97	-0.02	†	†	†	†
DECORATD	L	†	0.69	-2.58	0.14	13307	0.84	0.83	0.01	6506	0.87	0.87	0.00	1060	0.87	0.88	-0.01	615	0.86	0.90	-0.04	†	†	†	†
POURINT	L	†	0.78	-2.72	0.15	13327	0.89	0.88	0.01	6507	0.91	0.91	-0.01	1061	0.90	0.92	-0.03	612	0.87	0.93	-0.06	†	†	†	†
VEGETBLE	L , M	†	0.65	-1.51	0.12	16942	0.66	0.64	0.02	15405	0.75	0.75	0.00	3390	0.80	0.78	0.02	2980	0.74	0.81	-0.08	†	†	†	†
AWARDING	L , M	†	0.88	-0.84	0.27	16605	0.57	0.56	0.01	15213	0.70	0.69	0.01	3364	0.78	0.73	0.05	2948	0.75	0.77	-0.02	†	†	†	†
TRUNK	L , M	†	0.65	-1.15	0.00	16737	0.53	0.50	0.03	15311	0.65	0.64	0.01	3377	0.68	0.67	0.00	2967	0.63	0.72	-0.09	†	†	†	†
MOM	L , M	†	2.11	-0.74	0.00	16836	0.29	0.29	0.00	15380	0.60	0.60	0.01	3389	0.68	0.69	-0.01	2983	0.81	0.79	0.02	†	†	†	†
YELLOW	L , M	†	1.72	-0.63	0.00	16832	0.22	0.26	-0.04	15367	0.54	0.52	0.02	3392	0.70	0.61	0.09	2983	0.80	0.71	0.10	†	†	†	†
YOU	L , M	†	2.46	-0.38	0.00	16809	0.10	0.13	-0.03	15355	0.37	0.36	0.01	3390	0.54	0.47	0.08	2982	0.75	0.60	0.15	†	†	†	†
BOYBIRD	L , M , H	†	3.35	0.10	0.19	13255	0.21	0.23	-0.02	14548	0.35	0.37	-0.02	4387	0.48	0.47	0.01	15954	0.84	0.82	0.02	†	†	†	†
KAYLAFLY	L , M , H	†	0.59	-1.26	0.00	16928	0.54	0.53	0.01	15403	0.65	0.65	0.00	3391	0.69	0.68	0.01	2981	0.67	0.72	-0.06	†	†	†	†
COULDNOT	L , M , H	†	0.80	-1.32	0.00	16895	0.57	0.55	0.02	15381	0.70	0.71	-0.01	3389	0.77	0.75	0.02	2978	0.73	0.79	-0.06	†	†	†	†
COULD	L , M , H	†	0.54	-0.78	0.00	16884	0.44	0.42	0.02	15367	0.53	0.54	-0.01	3389	0.58	0.57	0.01	2982	0.58	0.61	-0.03	†	†	†	†
BEGWORD	M , H	†	0.78	-0.58	0.00	4274	0.64	0.58	0.06	12413	0.65	0.63	0.02	3988	0.70	0.68	0.03	15713	0.76	0.80	-0.04	†	†	†	†
QMARK	M , H	†	1.00	-0.54	0.00	4270	0.50	0.58	-0.08	12410	0.65	0.65	0.01	3987	0.76	0.70	0.06	15719	0.83	0.84	-0.01	†	†	†	†
TIME	M	†	0.90	-1.21	0.00	3610	0.77	0.76	0.01	8891	0.78	0.79	-0.01	2333	0.82	0.81	0.02	2368	0.82	0.83	-0.01	†	†	†	†
JOGGING	M	†	1.09	-0.80	0.12	3597	0.73	0.70	0.04	8890	0.73	0.74	-0.01	2328	0.79	0.76	0.03	2363	0.70	0.79	-0.09	†	†	†	†
ORSAT	M	†	2.48	-0.22	0.00	3606	0.31	0.31	0.00	8883	0.43	0.41	0.02	2329	0.52	0.47	0.05	2365	0.46	0.58	-0.12	†	†	†	†
ORPIG	M	†	1.89	-0.25	0.00	3612	0.41	0.37	0.05	8895	0.46	0.45	0.01	2329	0.49	0.50	-0.01	2365	0.47	0.58	-0.11	†	†	†	†
ORTAIL	M	†	2.79	-0.12	0.00	3611	0.28	0.22	0.06	8888	0.33	0.31	0.02	2329	0.36	0.37	-0.01	2361	0.38	0.48	-0.10	†	†	†	†
ORHAND	M	†	2.86	-0.04	0.00	3608	0.21	0.16	0.05	8888	0.28	0.24	0.03	2329	0.31	0.30	0.01	2363	0.28	0.41	-0.13	†	†	†	†
CATCH	M , H	†	3.37	0.18	0.00	4248	0.14	0.13	0.01	12369	0.22	0.23	-0.01	3983	0.31	0.34	-0.03	15714	0.77	0.75	0.02	†	†	†	†

See notes at end of table.

Table B1. Reading assessment item parameters and item fit by rounds: School years 1998–99, 1999–2000, and 2001–02—Continued

Reading	Test Form(s)	Test Form(s)	IRT parameters			Round 1				Round 2				Round 3				Round 4				Round 5			
							P+[5]				P+				P+				P+				P+		
	R1-R4	R5	a[1]	b[3]	c[3]	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference
FISHING	M , H	†	4.46	0.14	0.00	4239	0.12	0.12	0.00	12367	0.20	0.22	-0.02	3985	0.31	0.34	-0.03	15713	0.82	0.79	0.03	†	†	†	†
CANINBAG	M , H	†	1.89	0.15	0.22	2943	0.41	0.41	0.00	9590	0.49	0.49	0.00	3488	0.59	0.55	0.04	15390	0.79	0.78	0.01	†	†	†	†
KITNBED	M , H	†	2.85	0.16	0.16	3001	0.32	0.32	0.00	9162	0.41	0.42	-0.01	3291	0.52	0.51	0.01	15261	0.82	0.80	0.01	†	†	†	†
GIRLRED	M , H	†	1.41	0.14	0.00	3068	0.34	0.29	0.05	9647	0.36	0.38	-0.01	3482	0.45	0.45	0.00	15399	0.70	0.69	0.01	†	†	†	†
KIMCAD	M	†	4.41	0.36	0.50	894	0.52	0.51	0.01	1939	0.54	0.52	0.02	598	0.54	0.52	0.02	1149	0.57	0.55	0.02	†	†	†	†
NEEDHOME	M	†	4.87	0.02	0.16	796	0.23	0.24	-0.01	1790	0.29	0.31	-0.01	570	0.37	0.36	0.00	1139	0.61	0.50	0.11	†	†	†	†
LIKEDRY	M	†	3.92	0.49	0.25	784	0.31	0.26	0.05	1747	0.27	0.26	0.00	551	0.23	0.27	-0.03	1093	0.28	0.29	0.00	†	†	†	†
LIGHT	H	†	4.63	0.36	0.00	654	0.47	0.43	0.04	3515	0.38	0.40	-0.02	1655	0.45	0.49	-0.04	13347	0.75	0.75	0.00	†	†	†	†
KNOW	H	†	2.34	0.33	0.00	654	0.45	0.48	-0.03	3512	0.38	0.46	-0.08	1655	0.47	0.53	-0.06	13348	0.74	0.71	0.02	†	†	†	†
ELEPHANT	H	†	3.41	0.37	0.00	652	0.48	0.44	0.04	3505	0.41	0.41	0.00	1653	0.49	0.49	-0.01	13348	0.72	0.73	-0.01	†	†	†	†
WRONG	H	†	3.24	0.54	0.00	653	0.37	0.33	0.04	3510	0.34	0.29	0.05	1652	0.34	0.37	-0.03	13346	0.57	0.59	-0.01	†	†	†	†
ENVELOPE	H	†	3.25	0.78	0.00	649	0.28	0.20	0.08	3502	0.20	0.16	0.04	1654	0.24	0.22	0.02	13338	0.37	0.39	-0.02	†	†	†	†
DOGHOUSE	H	†	2.39	0.76	0.16	532	0.40	0.38	0.02	2569	0.35	0.36	-0.01	1301	0.42	0.41	0.01	12098	0.53	0.53	0.00	†	†	†	†
FLATTIRE	H	†	3.09	0.50	0.15	563	0.44	0.48	-0.04	2817	0.43	0.46	-0.03	1438	0.47	0.52	-0.04	12856	0.70	0.68	0.01	†	†	†	†
MARCHED	H	†	4.06	0.95	0.20	524	0.36	0.31	0.05	2531	0.31	0.29	0.02	1266	0.31	0.34	-0.03	11576	0.42	0.43	0.00	†	†	†	†
CHOCCAKE	H	†	4.52	0.66	0.17	530	0.45	0.41	0.04	2582	0.37	0.37	0.00	1311	0.42	0.44	-0.02	12163	0.59	0.59	0.00	†	†	†	†
RECIPE	H	†	3.24	1.16	0.19	509	0.34	0.25	0.09	2403	0.30	0.25	0.05	1223	0.30	0.27	0.03	11750	0.30	0.32	-0.03	†	†	†	†
INGREDNT	H	†	4.20	1.25	0.19	506	0.22	0.22	0.00	2397	0.25	0.22	0.03	1238	0.25	0.24	0.01	11410	0.25	0.27	-0.02	†	†	†	†
CATNAME	Hsup	†	2.28	0.95	0.00	†	†	†	†	†	†	†	†	1506	0.22	0.19	0.03	13309	0.30	0.30	0.00	†	†	†	†
OWNRNAME	Hsup	†	2.27	1.03	0.00	†	†	†	†	†	†	†	†	1525	0.16	0.16	0.00	13314	0.25	0.25	0.00	†	†	†	†
APPROX	Hsup	†	2.20	1.38	0.00	†	†	†	†	†	†	†	†	1519	0.07	0.07	0.00	13310	0.11	0.11	0.00	†	†	†	†
MOREINFO	Hsup	†	1.67	1.34	0.00	†	†	†	†	†	†	†	†	1500	0.09	0.11	-0.02	13306	0.17	0.16	0.00	†	†	†	†
UNUSUAL	Hsup	†	4.13	1.11	0.00	†	†	†	†	†	†	†	†	548	0.28	0.22	0.06	6180	0.30	0.30	0.00	†	†	†	†
WAGES	Hsup	†	1.72	1.62	0.00	†	†	†	†	†	†	†	†	308	0.18	0.14	0.04	3263	0.20	0.20	0.00	†	†	†	†
VICIOUS	Hsup	†	3.17	1.47	0.00	†	†	†	†	†	†	†	†	257	0.18	0.13	0.05	2509	0.20	0.20	0.00	†	†	†	†
MYSTERLY	Hsup	†	3.01	1.35	0.00	†	†	†	†	†	†	†	†	147	0.22	0.20	0.01	727	0.44	0.44	0.01	†	†	†	†
ALIGNMNT	Hsup	†	1.00	2.71	0.00	†	†	†	†	†	†	†	†	129	0.05	0.05	0.00	438	0.11	0.10	0.01	†	†	†	†
MAKE	Hsup	†	1.14	0.25	0.17	†	†	†	†	†	†	†	†	187	0.79	0.74	0.04	356	0.91	0.92	-0.01	†	†	†	†
MAINIDEA	Hsup	†	1.70	1.46	0.30	†	†	†	†	†	†	†	†	154	0.47	0.42	0.05	331	0.66	0.64	0.03	†	†	†	†
WHYNO	Hsup	†	1.14	1.36	0.00	†	†	†	†	†	†	†	†	161	0.24	0.26	-0.02	329	0.60	0.54	0.07	†	†	†	†
DESCRIBE	Hsup	†	2.28	1.80	0.12	†	†	†	†	†	†	†	†	170	0.15	0.17	-0.02	351	0.34	0.32	0.01	†	†	†	†
RDKNIGHT	†	R	2.12	1.21	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14274	0.56	0.56	0.00
RDSTORY	†	R	2.07	1.12	0.16	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14175	0.68	0.68	0.00
RDWAY	†	R	1.65	1.22	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14276	0.54	0.54	0.00
RDLIKE	†	R	1.24	0.94	0.11	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14111	0.70	0.70	0.00
RDTIME	†	R	2.15	0.69	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14148	0.85	0.85	0.00
RDDOMEST	†	R	2.01	1.57	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14277	0.34	0.33	0.00
RDGEORGR	†	L	2.48	0.99	0.21	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	3309	0.54	0.53	0.01
RDFEELSR	†	L	3.21	0.95	0.17	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	3298	0.53	0.53	0.00
RDGROWSR	†	L	1.76	0.80	0.14	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	3316	0.62	0.61	0.01
RDTANZAR	†	L	2.85	1.03	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	3511	0.34	0.34	0.01
RDLETR	†	L , M	2.42	0.81	0.26	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11253	0.84	0.84	0.00
RDDOCR	†	L , M	2.28	1.09	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11531	0.58	0.59	0.00
RDSAFER	†	L , M	2.30	1.44	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11527	0.34	0.33	0.00
RDSUPRIR	†	L , M	2.29	1.30	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11529	0.43	0.43	0.00

See notes at end of table.

Table B1. Reading assessment item parameters and item fit by rounds: School years 1998–99, 1999–2000, and 2001–02—Continued

Reading	Test Form(s) R1-R4	Test Form(s) R5	IRT parameters a[1]	b[2]	c[3]	Round 1 N	Round 1 P+[5] Actual	Round 1 P+ Predicted	Round 1 Difference	Round 2 N	Round 2 P+ Actual	Round 2 P+ Predicted	Round 2 Difference	Round 3 N	Round 3 P+ Actual	Round 3 P+ Predicted	Round 3 Difference	Round 4 N	Round 4 P+ Actual	Round 4 P+ Predicted	Round 4 Difference	Round 5 N	Round 5 P+ Actual	Round 5 P+ Predicted	Round 5 Difference
RDENDR	†	L , M	2.72	0.93	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11530	0.71	0.71	0.00
RDLIKER	†	L , M	2.20	1.56	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11528	0.26	0.26	0.00
RDSISR	†	L , M	2.69	1.12	0.15	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11209	0.64	0.65	0.00
RDDIFFR	†	L , M	1.48	1.50	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11525	0.35	0.34	0.01
RDCLUER	†	L , M	2.81	1.26	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11521	0.46	0.46	0.00
RDSAMER	†	L , M	2.16	0.99	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11521	0.65	0.65	-0.01
RDJAMEDR	†	L , M	1.81	1.23	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	11521	0.49	0.49	0.00
RDBOWY	†	M	2.69	1.30	0.19	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	7955	0.64	0.63	0.00
RDTRAINY	†	M	3.25	1.27	0.11	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	7996	0.63	0.63	0.00
RDSTRAGY	†	M	1.48	1.17	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	8029	0.60	0.60	0.00
RDFRICTY	†	M	1.63	1.37	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	8028	0.49	0.49	0.01
RDEMBOLY	†	M	0.81	3.63	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	8029	0.04	0.04	0.00
RDBEARY	†	M , H	2.23	0.99	0.08	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10710	0.82	0.82	-0.01
RDFINGRY	†	M , H	2.09	0.81	0.10	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10681	0.88	0.89	0.00
RDBIGKY	†	M , H	1.65	0.78	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10697	0.84	0.84	0.00
RDKINDY	†	M , H	1.58	1.09	0.10	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10690	0.73	0.73	0.00
RDAPOSTY	†	M , H	1.62	1.54	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10734	0.44	0.44	0.00
RDPOUCHY	†	M , H	2.60	1.50	0.05	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10539	0.48	0.48	0.00
RDFACTY	†	M , H	2.73	1.07	0.18	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10704	0.82	0.83	-0.01
RDBABONY	†	M , H	1.46	1.56	0.08	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10673	0.48	0.48	0.00
RDTRUEY	†	M , H	2.16	1.15	0.10	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10690	0.74	0.74	0.00
RDMALEBY	†	M , H	2.14	1.51	0.03	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10635	0.47	0.47	0.00
RDBSITY	†	M , H	3.20	1.12	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10717	0.77	0.78	-0.01
RDDANGRY	†	M , H	1.41	0.92	0.08	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10674	0.78	0.78	0.00
RDMOVEBY	†	M , H	1.43	1.42	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10651	0.51	0.51	0.00
RDJAKEY	†	M , H	3.37	1.41	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10713	0.52	0.52	0.00
RDBRETY	†	M , H	3.69	1.52	0.27	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10665	0.57	0.57	0.00
RDPROBLY	†	M , H	2.42	1.53	0.10	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10626	0.49	0.49	0.00
RDSOLVEY	†	M , H	2.78	1.68	0.06	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10636	0.34	0.34	0.00
RDHELPY	†	M , H	2.12	1.74	0.04	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10639	0.31	0.31	0.00
RDTEARB	†	H	2.20	1.51	0.14	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2711	0.70	0.70	0.00
RDPNEUMB	†	H	0.88	2.26	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2714	0.31	0.31	0.00
RDCROPB	†	H	1.94	2.18	0.15	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2690	0.33	0.32	0.01
RDCOMPRB	†	H	1.97	1.81	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2709	0.43	0.42	0.01
DCIRCLB	†	H	1.68	2.13	0.07	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2688	0.30	0.30	0.01
RDHOAXB	†	H	2.08	2.17	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2711	0.20	0.19	0.01
RDGUESS	†	H	1.65	2.10	0.21	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2590	0.44	0.42	0.01
RDMICROB	†	H	1.92	1.76	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2710	0.47	0.46	0.00
RDVORTXB	†	H	2.56	2.39	0.23	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2596	0.29	0.29	0.01
RDBAKEDB	†	H	1.26	1.49	0.17	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2690	0.68	0.67	0.00
RDANOMAB	†	H	0.45	3.59	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2393	0.22	0.19	0.03
RDMARIAB	†	H	1.20	0.70	0.33	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2705	0.92	0.92	0.00
RDJOSHB	†	H	2.01	1.54	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2710	0.62	0.62	0.00
RDRACHLB	†	H	2.57	1.59	0.16	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2695	0.67	0.67	0.00
RDPERSNB	†	H	2.40	1.74	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2710	0.48	0.47	0.00

See notes at end of table.

APPENDIX B

Table B1. Reading assessment item parameters and item fit by rounds: School years 1998–99, 1999–2000, and 2001–02—Continued

Reading	Test Form(s) R1-R4	Test Form(s) R5	IRT parameters a[1]	b[2]	c[3]	Round 1 N	Round 1 P+[5] Actual	Round 1 P+ Predicted	Round 1 Difference	Round 2 N	Round 2 P+ Actual	Round 2 P+ Predicted	Round 2 Difference	Round 3 N	Round 3 P+ Actual	Round 3 P+ Predicted	Round 3 Difference	Round 4 N	Round 4 P+ Actual	Round 4 P+ Predicted	Round 4 Difference	Round 5 N	Round 5 P+ Actual	Round 5 P+ Predicted	Round 5 Difference
RDWAGON	†	H	2.68	2.29	0.26	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2688	0.34	0.33	0.01
RDTHREEB	†	H	2.36	1.53	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2711	0.64	0.64	0.00
RDTHEMEB	†	H	2.82	1.64	0.12	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	2693	0.61	0.61	0.00
RUNS	R	R	3.05	-0.07	0.00	11573	0.10	0.10	-0.01	17489	0.29	0.30	-0.01	4823	0.44	0.44	-0.01	16186	0.87	0.84	0.03	14286	0.99	0.99	-0.01
DOWN	R	L	3.61	0.09	0.00	11581	0.06	0.06	0.00	17490	0.18	0.20	-0.02	4825	0.33	0.33	-0.01	16184	0.82	0.78	0.03	3540	0.94	0.95	-0.01
WENT	R	R	2.95	0.04	0.00	11574	0.08	0.08	0.00	17480	0.24	0.24	0.00	4820	0.36	0.37	-0.01	16185	0.81	0.79	0.02	14286	0.98	0.99	-0.01
JEEP	R	L	2.78	0.11	0.00	11576	0.08	0.07	0.01	17474	0.21	0.21	0.00	4827	0.32	0.34	-0.01	16184	0.77	0.75	0.02	3539	0.87	0.93	-0.06
BACKPACK	R	R	2.55	0.45	0.14	700	0.47	0.45	0.02	3353	0.47	0.45	0.02	1597	0.54	0.52	0.02	13463	0.68	0.68	-0.01	14256	0.95	0.94	0.01
RIDEBIKE	R	R	3.09	0.60	0.18	697	0.47	0.40	0.07	3283	0.42	0.39	0.03	1584	0.45	0.45	-0.01	13534	0.60	0.61	-0.01	14265	0.92	0.93	0.00
LISTEN	R	R	3.39	0.52	0.11	675	0.42	0.38	0.04	3090	0.39	0.38	0.01	1479	0.46	0.46	0.00	13191	0.64	0.64	0.00	14234	0.94	0.94	0.00
SIZES	R	R	3.87	0.64	0.13	662	0.36	0.33	0.03	2980	0.35	0.33	0.03	1442	0.40	0.39	0.00	12576	0.56	0.57	-0.01	14171	0.93	0.93	0.00
THROUGH	H	L	2.94	0.81	0.00	653	0.25	0.19	0.06	3504	0.17	0.16	0.01	1651	0.23	0.22	0.01	13346	0.38	0.37	0.00	3538	0.51	0.54	-0.03
RAGE	H	L	3.04	0.93	0.00	653	0.19	0.14	0.05	3510	0.12	0.11	0.01	1654	0.15	0.16	-0.01	13348	0.27	0.28	-0.01	3538	0.45	0.42	0.03
TOIL	H	L	2.18	0.99	0.00	653	0.21	0.15	0.06	3509	0.14	0.13	0.02	1655	0.17	0.17	0.00	13344	0.26	0.28	-0.02	3539	0.43	0.39	0.04
CAPTURE	H	L	2.55	1.06	0.00	643	0.15	0.11	0.04	3466	0.10	0.09	0.01	1644	0.11	0.13	-0.02	13287	0.22	0.23	0.00	3534	0.34	0.33	0.01
CORNER	H	L	2.27	1.01	0.00	645	0.17	0.14	0.03	3466	0.12	0.12	0.00	1644	0.15	0.16	-0.01	13282	0.25	0.26	-0.01	3536	0.41	0.37	0.04
WEB	H	L	1.78	1.09	0.00	645	0.18	0.14	0.03	3464	0.12	0.13	-0.01	1643	0.15	0.16	-0.01	13273	0.26	0.26	0.00	3535	0.36	0.34	0.02
STRANDS	H	L	1.57	1.57	0.00	645	0.08	0.06	0.03	3461	0.06	0.05	0.01	1643	0.06	0.07	-0.01	13268	0.11	0.11	0.00	3534	0.14	0.14	-0.01
QUIET	Hsup	M	2.99	0.76	0.00	†	†	†	†	†	†	†	†	1509	0.26	0.25	0.01	13334	0.41	0.41	0.00	8030	0.91	0.93	-0.01
WTLESS	Hsup	M	4.79	0.97	0.00	†	†	†	†	†	†	†	†	1487	0.13	0.13	0.00	13324	0.22	0.23	-0.01	8030	0.90	0.89	0.01
REQUIRE	Hsup	M	3.75	1.05	0.00	†	†	†	†	†	†	†	†	1456	0.14	0.11	0.02	13330	0.19	0.19	0.00	8029	0.80	0.81	0.00
MOISTURE	Hsup	M	2.71	1.13	0.00	†	†	†	†	†	†	†	†	358	0.32	0.32	0.00	3929	0.40	0.42	-0.02	8030	0.71	0.70	0.01
PREFRNCE	Hsup	R	1.45	1.64	0.00	†	†	†	†	†	†	†	†	220	0.28	0.18	0.10	1858	0.33	0.26	0.07	14279	0.32	0.33	-0.01
AMBITIO	Hsup	R	2.18	1.77	0.00	†	†	†	†	†	†	†	†	179	0.11	0.08	0.03	1207	0.22	0.15	0.06	14278	0.21	0.21	0.00
CRITCISM	Hsup	R	2.59	1.65	0.00	†	†	†	†	†	†	†	†	158	0.11	0.10	0.02	905	0.15	0.21	-0.06	14277	0.27	0.26	0.01

† Not applicable.

[1] Parameter for discrimination.

[2] Parameter for difficulty.

[3] Parameter for guessing.

[4] Number in sample.

[5] Proportion correct.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. Not all items appeared in test forms for all rounds.

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

APPENDIX B

Table B2. Mathematics assessment item parameters and item fit by rounds: School year 1998–99, 1999–2000, and 2001–02

Mathematics	Test Form(s) R1-R4	Test Form(s) R5	IRT parameters			Round 1				Round 2				Round 3				Round 4				Round 5			
							P+[5]		Differ-		P+		Differ-		P+		Differ-		P+		Differ-		P+		Differ-
			a[1]	b[2]	c[3]	N[4]	Actual	Predicted	ence	N	Actual	Predicted	ence	N	Actual	Predicted	ence	N	Actual	Predicted	ence	N	Actual	Predicted	ence
SM-LG-SM	R	†	1.32	-0.97	0.23	18379	0.60	0.60	0.00	19471	0.80	0.79	0.01	5206	0.89	0.86	0.03	16587	0.94	0.95	-0.01	†	†	†	†
COUNT 20	R	†	1.13	-1.20	0.00	18622	0.56	0.57	-0.01	19651	0.80	0.77	0.03	5222	0.85	0.85	0.00	16645	0.92	0.94	-0.03	†	†	†	†
NUMBER 9	R	†	2.34	-1.34	0.00	18631	0.66	0.67	-0.01	19652	0.90	0.90	0.01	5226	0.95	0.95	0.00	16647	0.99	0.99	0.00	†	†	†	†
NUMBER23	R	†	1.89	-0.70	0.00	18614	0.33	0.34	-0.01	19639	0.65	0.64	0.01	5225	0.77	0.77	-0.01	16647	0.94	0.94	0.01	†	†	†	†
3RD LINE	R	†	1.83	-0.66	0.00	18636	0.33	0.32	0.01	19653	0.63	0.62	0.01	5225	0.76	0.75	0.01	16647	0.91	0.93	-0.02	†	†	†	†
STICKBAT	R	†	0.89	-1.64	0.06	18616	0.74	0.71	0.03	19647	0.84	0.84	-0.01	5224	0.89	0.89	0.00	16643	0.92	0.95	-0.03	†	†	†	†
_ 78910	R	†	1.78	-0.63	0.00	18621	0.28	0.31	-0.04	19649	0.63	0.60	0.02	5225	0.78	0.74	0.04	16644	0.93	0.92	0.01	†	†	†	†
3+2 CARS	R	†	1.27	-0.65	0.00	18636	0.38	0.35	0.03	19656	0.59	0.60	-0.01	5225	0.71	0.71	0.00	16646	0.86	0.88	-0.02	†	†	†	†
5-1ORANG	R	†	1.76	-0.10	0.12	18427	0.27	0.24	0.03	19524	0.43	0.43	0.00	5215	0.56	0.56	0.00	16585	0.79	0.81	-0.02	†	†	†	†
2CRAYONS	L	†	1.56	-3.01	0.00	14378	0.98	0.98	0.00	8443	0.99	0.99	0.00	1353	0.99	0.99	-0.01	1097	0.98	1.00	-0.01	†	†	†	†
3BANANAS	L	†	0.78	-2.71	0.08	14289	0.89	0.87	0.02	8412	0.91	0.92	-0.01	1350	0.92	0.92	-0.01	1092	0.87	0.94	-0.07	†	†	†	†
6BANANAS	L	†	1.10	-1.09	0.00	14364	0.44	0.44	0.00	8440	0.59	0.58	0.01	1352	0.66	0.61	0.05	1095	0.71	0.67	0.04	†	†	†	†
NUMBER 4	L	†	3.10	-1.90	0.00	14369	0.87	0.88	0.00	8441	0.96	0.97	-0.01	1353	0.96	0.97	-0.01	1097	0.97	0.98	-0.01	†	†	†	†
NUMBER 7	L	†	2.75	-1.62	0.00	14372	0.74	0.75	-0.01	8441	0.91	0.90	0.00	1353	0.90	0.92	-0.02	1096	0.93	0.95	-0.02	†	†	†	†
NUMBER17	L , M	†	1.89	-0.87	0.00	17477	0.37	0.39	-0.02	14603	0.66	0.63	0.03	2873	0.71	0.71	-0.01	3411	0.84	0.83	0.01	†	†	†	†
SQUARE	L	†	0.93	-2.51	0.15	14356	0.90	0.88	0.02	8427	0.91	0.93	-0.02	1350	0.90	0.93	-0.04	1097	0.90	0.95	-0.05	†	†	†	†
LG-SM-SM	L , M	†	1.46	-0.98	0.28	17287	0.61	0.61	0.00	14487	0.77	0.76	0.01	2862	0.83	0.81	0.02	3376	0.87	0.88	-0.01	†	†	†	†
000X	L , M	†	1.05	-0.79	0.19	16857	0.51	0.51	0.00	14133	0.67	0.64	0.03	2823	0.75	0.69	0.06	3327	0.76	0.77	-0.01	†	†	†	†
HALFOVAL	L , M	†	0.90	-0.63	0.22	16822	0.50	0.49	0.00	14113	0.63	0.61	0.03	2798	0.68	0.65	0.03	3320	0.73	0.72	0.01	†	†	†	†
2+3STICK	L , M , H	†	1.41	-0.57	0.00	18623	0.32	0.31	0.01	19650	0.56	0.57	-0.01	5224	0.71	0.69	0.01	16645	0.88	0.88	0.00	†	†	†	†
3-1PENCL	L , M	†	0.81	-1.64	0.00	17494	0.68	0.66	0.01	14609	0.78	0.78	0.00	2873	0.80	0.81	-0.01	3411	0.80	0.86	-0.06	†	†	†	†
2+5CIRCL	L , M , H	†	1.49	0.01	0.00	18623	0.15	0.13	0.02	19648	0.32	0.31	0.01	5221	0.46	0.45	0.01	16643	0.70	0.72	-0.03	†	†	†	†
8-6CRAYN	L , M , H	†	1.20	-0.24	0.00	18625	0.22	0.22	0.00	19651	0.42	0.43	-0.01	5222	0.53	0.55	-0.02	16644	0.80	0.77	0.02	†	†	†	†
PNTBRUSH	L , M , H	†	1.51	-1.24	0.21	18299	0.68	0.69	0.00	19480	0.87	0.86	0.01	5200	0.93	0.92	0.01	16622	0.98	0.98	0.00	†	†	†	†
#CHOC	L , M , H	†	1.26	-1.35	0.00	18609	0.64	0.63	0.00	19647	0.83	0.83	0.00	5222	0.90	0.89	0.01	16643	0.96	0.96	-0.01	†	†	†	†
#VANILLA	L , M , H	†	1.19	-1.55	0.00	18608	0.70	0.70	0.00	19647	0.86	0.87	0.00	5222	0.93	0.91	0.02	16643	0.97	0.97	-0.01	†	†	†	†
#BUGS	L , M , H	†	1.38	-0.53	0.22	18166	0.48	0.45	0.03	19399	0.65	0.65	0.00	5166	0.75	0.75	0.00	16564	0.88	0.90	-0.02	†	†	†	†
4LINES	M	†	0.57	-1.14	0.18	3032	0.76	0.73	0.03	6041	0.77	0.75	0.02	1499	0.80	0.76	0.04	2275	0.73	0.79	-0.06	†	†	†	†
SHAPES	M	†	0.65	0.02	0.22	3033	0.52	0.52	0.00	6049	0.55	0.54	0.01	1505	0.60	0.56	0.04	2294	0.59	0.60	-0.01	†	†	†	†
PATTERN	M , H	†	1.29	0.00	0.22	4243	0.50	0.51	-0.01	11185	0.59	0.59	0.00	3867	0.67	0.65	0.02	15533	0.79	0.80	-0.01	†	†	†	†
2 + 2	M	†	3.34	-0.47	0.00	3122	0.51	0.57	-0.05	6167	0.67	0.70	-0.02	1520	0.77	0.77	0.00	2317	0.89	0.88	0.02	†	†	†	†
3 + 4	M , H	†	2.04	-0.14	0.00	4258	0.34	0.40	-0.06	11200	0.54	0.55	0.00	3869	0.67	0.66	0.01	15547	0.86	0.86	0.00	†	†	†	†
1 + 7	M	†	1.31	-0.38	0.00	3121	0.48	0.49	0.00	6159	0.53	0.55	-0.02	1518	0.62	0.59	0.03	2315	0.72	0.67	0.04	†	†	†	†
3 + 3	M	†	4.27	-0.38	0.00	3120	0.43	0.46	-0.03	6165	0.60	0.62	-0.02	1519	0.71	0.72	-0.01	2317	0.86	0.86	0.00	†	†	†	†
11 + 3	M , H	†	2.03	0.24	0.00	4250	0.17	0.19	-0.02	11179	0.30	0.31	-0.01	3866	0.43	0.42	0.01	15542	0.70	0.69	0.01	†	†	†	†
12 + 6	M , H	†	1.68	0.50	0.00	4243	0.12	0.13	-0.01	11169	0.21	0.21	-0.01	3860	0.31	0.29	0.01	15539	0.54	0.54	0.01	†	†	†	†
17 - 4	M , H	†	2.31	0.69	0.00	4245	0.04	0.05	-0.01	11168	0.09	0.10	-0.01	3859	0.15	0.17	-0.02	15536	0.44	0.42	0.02	†	†	†	†
# STRAW	H	†	1.07	-1.78	0.00	1136	0.95	0.97	-0.02	5042	0.97	0.97	0.00	2351	0.98	0.98	0.01	13232	0.99	0.99	0.00	†	†	†	†
# MORE	H	†	1.74	0.53	0.00	1135	0.35	0.25	0.09	5038	0.36	0.31	0.04	2351	0.43	0.38	0.05	13221	0.54	0.58	-0.04	†	†	†	†
24-14BKS	H	†	2.45	1.00	0.00	1136	0.07	0.05	0.01	5038	0.09	0.08	0.01	2350	0.13	0.12	0.01	13229	0.26	0.27	-0.01	†	†	†	†
CHANGE	H	†	1.62	2.04	0.00	1135	0.04	0.01	0.04	5037	0.01	0.01	0.00	2351	0.02	0.02	0.00	13229	0.03	0.04	0.00	†	†	†	†
17CENTS	H	†	2.50	1.19	0.00	1134	0.02	0.03	0.00	5030	0.04	0.04	0.00	2350	0.05	0.06	-0.01	13221	0.17	0.17	0.01	†	†	†	†
BDCAKE	H	†	1.94	1.23	0.00	1135	0.04	0.04	-0.01	5036	0.05	0.06	-0.01	2347	0.07	0.08	-0.01	13223	0.19	0.19	0.01	†	†	†	†
24/4 TAB	H	†	1.42	1.51	0.00	1135	0.06	0.04	0.02	5034	0.08	0.06	0.02	2347	0.09	0.07	0.01	13217	0.14	0.14	-0.01	†	†	†	†

See notes at end of table.

Table B2. Mathematics assessment item parameters and item fit by rounds: School years 1998–99, 1999–2000, and 2001–02—Continued

Mathematics	Test Form(s) R1-R4	Test Form(s) R5	IRT parameters			Round 1				Round 2				Round 3				Round 4				Round 5			
							P+		Differ-		P+		Differ-		P+		Differ-		P+		Differ-		P+		Differ-
			a[1]	b[2]	c[3]	N	Actual	Predicted	ence	N	Actual	Predicted	ence	N	Actual	Predicted	ence	N	Actual	Predicted	ence	N	Actual	Predicted	ence
2-1+2	H	†	1.59	0.56	0.00	1134	0.28	0.25	0.03	5027	0.30	0.31	-0.02	2347	0.35	0.37	-0.02	13221	0.56	0.56	0.00	†	†	†	†
2-SEP	H	†	2.61	0.30	0.00	1135	0.25	0.34	-0.09	5035	0.35	0.43	-0.08	2347	0.43	0.52	-0.09	13230	0.80	0.76	0.04	†	†	†	†
3-JUL	H	†	2.49	0.27	0.00	1136	0.27	0.36	-0.09	5034	0.38	0.45	-0.08	2347	0.45	0.54	-0.09	13229	0.81	0.77	0.04	†	†	†	†
6+7	H	†	2.05	0.39	0.00	1130	0.28	0.30	-0.03	5022	0.36	0.38	-0.02	2346	0.45	0.46	-0.01	13225	0.68	0.68	0.00	†	†	†	†
9-DEC	H	†	2.27	0.71	0.00	1131	0.10	0.14	-0.04	5020	0.14	0.19	-0.05	2343	0.20	0.25	-0.05	13218	0.49	0.47	0.03	†	†	†	†
26 + 20	H	†	2.36	0.77	0.00	1131	0.10	0.11	-0.01	5000	0.13	0.15	-0.03	2339	0.17	0.21	-0.04	13210	0.44	0.42	0.02	†	†	†	†
COUNT10	R	†	0.82	-2.76	0.00	18622	0.90	0.90	0.01	19651	0.96	0.95	0.01	5222	0.96	0.97	0.00	16645	0.97	0.99	-0.01	†	†	†	†
TIME	†	R	1.85	1.08	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14378	0.63	0.63	0.00
NICKELS	†	R	2.14	1.16	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14372	0.59	0.59	0.00
PAPERS	†	R	2.47	1.13	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14376	0.62	0.62	0.00
SPOONS	†	R	2.42	1.33	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14373	0.50	0.49	0.00
FRIES	†	R	2.45	1.37	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14371	0.48	0.47	0.01
NUMBER	†	R	2.18	1.33	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14375	0.49	0.49	0.00
CHART	†	R	1.76	1.29	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14376	0.51	0.51	0.00
FRUIT	†	R	1.64	1.49	0.11	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14184	0.48	0.48	0.00
INCHES	†	R	1.65	1.76	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14371	0.28	0.28	0.00
TILES	†	R	1.32	1.67	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14369	0.34	0.34	0.00
PAIRS	†	R	2.63	1.75	0.12	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	13786	0.33	0.33	0.00
MARBLES	†	R	2.19	2.11	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	14360	0.12	0.11	0.00
ORANGE_R	†	L	1.54	0.22	0.15	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4216	0.82	0.82	0.00
PATHS_R	†	L	1.02	0.31	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4209	0.69	0.68	0.01
VICKS_R	†	L	2.13	-0.26	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4224	0.96	0.96	0.00
TEAMS_R	†	L	0.99	-0.38	0.15	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4125	0.89	0.88	0.01
MONEY_R	†	L	3.03	1.06	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4221	0.26	0.24	0.01
SQUARE_R	†	L	0.84	0.84	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4223	0.47	0.47	0.00
BEADS_R	†	L	3.57	0.97	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4221	0.33	0.31	0.02
RULER_R	†	L	1.13	0.64	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4223	0.56	0.56	0.01
PAGES_R	†	L	2.09	0.69	0.20	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4084	0.67	0.65	0.01
SIDES_R	†	L , M	1.41	0.67	0.11	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	9527	0.73	0.73	0.00
NEXT_R	†	L	2.96	1.26	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4220	0.14	0.12	0.02
POINTS_R	†	L , M	1.63	1.10	0.29	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	9417	0.65	0.64	0.01
MEANS_R	†	L , M	2.49	1.05	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	9557	0.52	0.53	0.00
EQUAL_R	†	L , M	2.61	1.05	0.17	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	9412	0.62	0.61	0.01
NUMBE1_R	†	L , M	2.95	1.22	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	9554	0.40	0.40	0.00
DO_Y	†	M	2.13	1.04	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5338	0.71	0.71	0.00
CUBE_Y	†	M	1.08	1.26	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5339	0.52	0.52	0.00
MOST_Y	†	M	2.35	0.55	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5339	0.94	0.94	0.00
FEWEST_Y	†	M	2.18	0.83	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5339	0.83	0.84	-0.01
MORE1_Y	†	M	3.18	1.21	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5339	0.61	0.61	0.00
FEWER_Y	†	M	3.16	1.28	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5339	0.53	0.53	0.00
NEXT_Y	†	M	2.03	0.96	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5340	0.75	0.75	-0.01
SCORE_Y	†	M	2.75	1.11	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5340	0.68	0.69	-0.01
DESIGN_Y	†	M	1.47	1.48	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5336	0.40	0.40	0.00
STAR-Y	†	M	1.13	1.40	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	5341	0.46	0.46	0.00
BOX2-Y	†	M , H	2.92	1.32	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10142	0.67	0.67	0.00

See notes at end of table.

Table B2. Mathematics assessment item parameters and item fit by rounds: School years 1998–99, 1999–2000, and 2001–02—Continued

Mathematics	Test Form(s) R1-R4	Test Form(s) R5	IRT parameters			Round 1				Round 2				Round 3				Round 4				Round 5			
							P+		Difference		P+		Difference		P+		Difference		P+		Difference		P+		Difference
			a[1]	b[2]	c[3]	N	Actual	Predicted		N	Actual	Predicted		N	Actual	Predicted		N	Actual	Predicted		N	Actual	Predicted	
CHILDR_Y	†	M , H	2.29	1.38	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10138	0.60	0.61	0.00
PAGES_Y	†	M , H	2.31	1.37	0.06	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	9988	0.64	0.64	0.00
MORE2_Y	†	M , H	2.41	1.44	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10128	0.57	0.57	0.00
CHARGE_Y	†	M , H	1.60	1.66	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10136	0.42	0.42	0.00
PENCIL_Y	†	M , H	1.05	1.81	0.05	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10126	0.41	0.41	0.00
SECOND_Y	†	M , H	2.19	1.42	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10139	0.57	0.57	0.00
MINUTE_Y	†	M , H	2.38	1.83	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	10137	0.28	0.28	0.00
MARIA_B	†	H	2.52	1.58	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4802	0.66	0.66	-0.01
LOUISA_B	†	H	2.55	1.78	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4799	0.49	0.49	0.00
CARDS_B	†	H	1.69	1.66	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4799	0.58	0.58	0.00
EQUAL_B	†	H	1.71	1.85	0.09	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4724	0.51	0.51	0.00
LARGER_B	†	H	1.57	1.77	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4803	0.51	0.51	0.00
NUMBE2_B	†	H	1.82	1.96	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4800	0.38	0.38	0.01
AREA_B	†	H	1.47	1.73	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4802	0.53	0.53	0.00
MONEY_B	†	H	1.74	2.05	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4800	0.33	0.33	0.00
EDGES_B	†	H	1.00	2.20	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4801	0.34	0.34	0.00
FENCE_B	†	H	1.90	2.52	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4789	0.11	0.11	0.00
MYSTER_B	†	H	2.18	2.11	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4801	0.27	0.26	0.00
LABEL-B	†	H	1.56	2.24	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4798	0.25	0.25	0.00
SAME_B	†	H	1.39	2.44	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4798	0.19	0.19	0.00
TILES_B	†	H	2.15	2.88	0.00	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4797	0.03	0.03	0.00
TALL_B	†	H	2.24	2.05	0.04	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	4621	0.33	0.33	0.00
51015_25	R	L	2.06	0.20	0.00	18618	0.04	0.06	-0.02	19636	0.19	0.20	-0.01	5224	0.30	0.34	-0.04	16644	0.71	0.67	0.04	4229	0.90	0.84	0.07
2+5MARBL	R	L	1.26	-0.09	0.00	18616	0.19	0.17	0.01	19645	0.38	0.37	0.01	5223	0.53	0.49	0.04	16645	0.75	0.74	0.01	4229	0.69	0.85	-0.16
3+7PENNY	R	L	1.83	0.10	0.00	18623	0.10	0.09	0.02	19645	0.24	0.25	-0.01	5224	0.38	0.40	-0.01	16645	0.72	0.71	0.01	4229	0.85	0.86	-0.01
13_79	R	R	1.62	0.63	0.00	4259	0.09	0.10	-0.02	11212	0.14	0.17	-0.03	3872	0.21	0.24	-0.04	15550	0.50	0.46	0.04	14380	0.81	0.81	0.00
COST$10	R	R	2.02	0.69	0.00	4173	0.11	0.06	0.04	11191	0.14	0.12	0.02	3866	0.22	0.19	0.04	15546	0.39	0.42	-0.03	14380	0.82	0.81	0.00
8/2CANDY	R	R	2.03	0.93	0.00	4260	0.05	0.03	0.02	11204	0.08	0.07	0.02	3872	0.15	0.11	0.04	15548	0.26	0.29	-0.03	14380	0.72	0.71	0.01
15/5CARS	R	R	2.09	0.78	0.00	4260	0.08	0.05	0.03	11201	0.11	0.09	0.02	3871	0.17	0.15	0.02	15542	0.35	0.37	-0.02	14380	0.78	0.78	0.00
12 BY 2S	M , H	L	1.84	0.03	0.00	4253	0.31	0.31	0.00	11199	0.43	0.44	-0.01	3863	0.52	0.55	-0.03	15546	0.79	0.78	0.01	4226	0.88	0.88	0.00
HEADSUP	H	L	1.08	0.98	0.00	1136	0.23	0.18	0.05	5040	0.22	0.22	0.00	2351	0.26	0.25	0.00	13225	0.35	0.37	-0.03	4220	0.47	0.41	0.07
HOWMANY$	H	L	1.41	0.91	0.00	1135	0.19	0.15	0.04	5035	0.20	0.19	0.02	2351	0.25	0.23	0.02	13229	0.34	0.38	-0.04	4226	0.50	0.42	0.08
12-? PEN	H	L	2.26	0.97	0.00	1136	0.12	0.07	0.06	5040	0.13	0.09	0.04	2351	0.16	0.14	0.02	13226	0.27	0.30	-0.03	4220	0.38	0.35	0.03
GOALS	H	R	2.00	1.12	0.00	1135	0.05	0.05	0.00	5036	0.07	0.07	0.00	2351	0.10	0.11	0.00	13226	0.23	0.23	0.00	14378	0.62	0.61	0.00
4+4-2	H	L	1.99	0.67	0.00	1131	0.16	0.17	-0.01	5025	0.23	0.23	0.00	2345	0.30	0.29	0.02	13222	0.51	0.50	0.01	4227	0.53	0.57	-0.04

† Not applicable.

[1] Parameter for discrimination.

[2] Parameter for difficulty.

[3] Parameter for guessing.

[4] Number in sample.

[5] Proportion correct.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third grade have been put on a common scale to support comparisons. Not all items appeared in test forms for all rounds.

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

Table B3. Science assessment item parameters and item fit by rounds: School years 1998–99, 1999–2000, and 2001–02

Science	Test Form(s)	IRT parameters			Round 5			
						P+[5]		
	R5	a[1]	b[2]	c[3]	N[4]	Actual	Predicted	Difference
ROUIMM	R	1.17	-1.45	0.04	14274	0.85	0.85	0.00
ROUJUN	R	1.42	-0.80	0.00	14354	0.68	0.68	0.00
ROUBRN	R	1.52	-0.67	0.00	14355	0.64	0.64	0.00
ROUFRZ	R	1.60	-0.97	0.12	14241	0.78	0.79	0.00
ROUERT	R	0.95	-0.72	0.12	14300	0.67	0.67	0.00
ROUTAP	R	0.76	-0.65	0.01	14325	0.59	0.59	0.00
ROUJAR	R	0.60	-0.60	0.00	14353	0.56	0.56	0.00
ROUSRF	R	1.16	-0.40	0.42	14265	0.73	0.73	0.00
ROUSHD	R	1.03	0.13	0.00	14348	0.35	0.35	0.00
ROUMCE	R	1.56	0.16	0.26	14101	0.49	0.48	0.01
ROUMTN	R	1.46	0.21	0.21	14125	0.44	0.44	0.01
ROUFLY	R	1.40	0.21	0.14	14159	0.39	0.39	0.01
ROUBLB	R	0.91	0.27	0.22	14130	0.47	0.47	0.00
ROUGRT	R	1.19	0.18	0.08	14266	0.37	0.37	0.00
ROUSOL	R	0.67	0.15	0.15	14269	0.47	0.47	0.00
RENRGY	L	1.32	-1.89	0.16	4172	0.85	0.85	0.00
RPLANT	L	1.51	-1.83	0.16	4191	0.85	0.85	-0.01
RBULB	L	0.93	-1.92	0.14	4140	0.80	0.80	0.00
RDSAST	L	1.53	-1.47	0.08	4180	0.71	0.72	0.00
RORGAN	L	0.59	-1.62	0.14	4126	0.68	0.68	0.01
ROCCUR	L	1.62	-0.94	0.21	4114	0.56	0.55	0.01
RFGRPS	L	0.62	-1.39	0.22	4122	0.68	0.67	0.01
RANIML	L	0.97	-1.12	0.09	4141	0.56	0.56	0.00
RTOOL	L	0.94	-1.55	0.11	4181	0.70	0.70	0.00
RSUNIS	L	1.22	-0.72	0.28	4171	0.53	0.53	0.01
RWINGS	L	1.86	-1.20	0.10	4169	0.61	0.61	0.00
RWATER	L	1.13	-0.76	0.10	4114	0.44	0.43	0.01
RFISHB	L	1.07	-0.64	0.10	4144	0.40	0.40	0.01
RPWDER	L	1.04	-0.59	0.15	3978	0.43	0.41	0.01
RTHING	L , M	1.22	-0.76	0.21	11279	0.68	0.68	0.00
RSEEDS	L , M	0.94	-0.85	0.07	11327	0.63	0.63	0.00
RDESRT	L , M	1.01	-0.47	0.15	11261	0.56	0.56	0.00
RHEART	L , M	1.37	-0.60	0.00	11396	0.53	0.53	0.00
RFORMS	L , M	0.98	-1.36	0.16	11230	0.80	0.80	0.00
RSHAPE	L , M	1.15	-0.63	0.18	11188	0.62	0.62	0.00
YDSAST	M	0.83	-0.74	0.14	7174	0.70	0.70	0.00
YMOON	M	1.51	-0.25	0.30	7124	0.65	0.65	0.01
YTHEMT	M	0.77	-0.40	0.13	7130	0.61	0.60	0.00
YINSCT	M	1.27	-0.26	0.18	7183	0.60	0.59	0.00
YSENSE	M	1.04	-0.11	0.00	7200	0.45	0.44	0.00
YSOUND	M	0.98	-0.27	0.09	7162	0.55	0.55	0.00
YBLANC	M , H	1.12	0.53	0.03	10156	0.31	0.31	0.01
YBEES	M	1.04	0.08	0.00	7200	0.37	0.37	0.01

See notes at end of table.

Table B3. Science assessment item parameters and item fit by rounds: School years 1998–99, 1999–2000, and 2001–02—Continued

Science	Test Form(s) R5	IRT parameters			Round 5			
						P+		
		a[1]	b[2]	c[3]	N	Actual	Predicted	Difference
YDSOLV	M , H	0.79	0.21	0.12	10109	0.49	0.49	0.00
YPLAIN	M , H	0.79	-1.25	0.00	10151	0.82	0.82	0.00
YFARMG	M , H	0.52	0.57	0.00	10148	0.38	0.38	0.00
YFWATE	M , H	1.39	0.33	0.25	10148	0.50	0.50	0.01
YLIVE	M , H	1.45	0.45	0.12	10137	0.37	0.37	0.01
YHUMID	M , H	0.73	1.20	0.10	10111	0.28	0.27	0.00
BSHADW	H	1.16	-0.03	0.14	2951	0.75	0.76	-0.01
BPLANT	H	1.33	0.31	0.13	2946	0.64	0.64	0.00
BEARTH	H	0.78	-0.33	0.00	2952	0.73	0.74	0.00
BPLNT2	H	0.98	0.45	0.11	2927	0.57	0.57	0.00
BSOUND	H	1.22	0.50	0.12	2930	0.56	0.56	0.00
BHIBER	H	0.75	0.54	0.21	2933	0.60	0.59	0.00
BSLIDE	H	1.41	0.86	0.11	2940	0.39	0.39	0.01
BSTORM	H	1.71	1.05	0.18	2882	0.36	0.35	0.01
BPLLUT	H	0.84	0.94	0.16	2939	0.46	0.46	0.00
BMAMML	H	0.86	1.20	0.00	2949	0.28	0.28	0.00
BPOLAR	H	1.22	1.02	0.09	2938	0.34	0.34	0.01
BPLNT3	H	1.34	1.20	0.06	2948	0.25	0.24	0.01
BSOIL	H	1.04	1.29	0.13	2935	0.32	0.32	0.01

[1] Parameter for discrimination.

[2] Parameter for difficulty.

[3] Parameter for guessing.

[4] Number in sample.

[5] Proportion correct.

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0). Estimates for kindergarten through third 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), spring 2002.

This page is intentionally left blank.