U.S. Department of Education
Institute of Education Sciences
NCES 2006–036

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

Psychometric Report for the Fifth Grade

November 2005

Judith M. Pollack
Michelle Najarian
Donald A. Rock
Educational Testing Service

Sally Atkins-Burnett
University of Toledo

Elvira Germino Hausken
Project Officer
National Center for Education Statistics

www.manaraa.com

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Statistics
Mark Schneider
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

November 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 address shown above.

Suggested Citation

Pollack, J.M., Atkins-Burnett, S., Najarian, M., and Rock, D.A. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for the Fifth Grade* (NCES 2006–036). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Content Contact:

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

www.manaraa.com

المنارة للاستشارات
www.manaraa.com

TABLE OF CONTENTS

www.manaraa.com

TABLE OF CONTENTS (continued)

www.manaraa.com

TABLE OF CONTENTS (continued)

www.manaraa.com

TABLE OF CONTENTS (continued)

List of Appendixes

www.manaraa.com

TABLE OF CONTENTS (continued)

List of Tables

www.manaraa.com

TABLE OF CONTENTS (continued)

List of Tables (continued)

www.manaraa.com

TABLE OF CONTENTS (continued)

List of Tables (continued)

www.manaraa.com

TABLE OF CONTENTS (continued)

List of Tables (continued)

www.manaraa.com

TABLE OF CONTENTS (continued)

List of Tables (continued)

www.manaraa.com

TABLE OF CONTENTS (continued)

List of Tables (continued)

www.manaraa.com

TABLE OF CONTENTS (continued)

Appendix A Tables (continued)

www.manaraa.com

Appendix A Tables (continued)

www.manaraa.com

TABLE OF CONTENTS (continued)

Appendix A Tables (continued)

Appendix B Tables

www.manaraa.com

TABLE OF CONTENTS (continued)

Appendix C Tables

Appendix D Tables

www.manaraa.com

TABLE OF CONTENTS (continued)

List of Figures

www.manaraa.com

TABLE OF CONTENTS (continued)

List of Exhibits

www.manaraa.com

1. INTRODUCTION

This report documents the design, construction, and psychometric characteristics of the assessment instruments used in the spring 2004 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.

While the ECLS-K spans kindergarten through fifth grade, this report documents the psychometric results for the sixth round of data collection, in spring 2004, when approximately 90 percent of the sampled children were in fifth grade. Also included is a review of the salient features of the assessments used in kindergarten through third grade. Among these salient features are the selection and design of assessment instruments and selected psychometric characteristics.

Two domains are represented by the ECLS-K fifth-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 fifth 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 fifth grade were designed to measure an individual child's knowledge at a given point in time, as well as that same child's academic growth in each subject on vertical score scales based on successive assessments. The score scales for reading and mathematics

www.manaraa.com

measure growth from fall-kindergarten through fifth grade, while the science assessment was administered only in the third- and fifth-grade rounds.

The cognitive assessments were designed not only to make reliable normative comparisons with respect to status and growth, but also to provide criterion-referenced interpretations. That is, in the reading and mathematics content domains, criterion-referenced proficiency scores can be used to describe a given child's mastery of specific knowledge and skills that mark ascending critical points on the developmental growth curve. These multiple criterion-referenced levels serve two functions. First, they help 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 relationship of certain school processes to 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 fifth-grade assessment instruments. Differences between the kindergarten-first grade (K-1), third-grade, and fifth-grade assessment batteries are described. For the direct cognitive tests, chapter 2 includes selection of content domains, notes on frameworks, descriptions of field testing, and selection of test items. It describes the criterion-referenced subsets of items in the reading and mathematics tests that were used to mark proficiency levels in kindergarten through third grade and the extension of these levels for fifth-grade skills. Chapter 2 also describes the evaluation of potential gaps in the longitudinal scale for the years in which data were not collected, second and fourth grades, and the steps taken to avoid compromising measurement of gains. For the indirect measures, chapter 2 describes the development and content of the

www.manaraa.com

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 a description of the quality control procedures applied to analysis of the assessment data, as well as an overview of item response theory (IRT) procedures used in computing test scores and the differential item functioning (DIF) procedures used to detect problem items. Chapter 4 presents the psychometric characteristics of the direct cognitive tests given in fifth grade, and chapter 5 describes their role in longitudinal measurement. Chapter 6 describes the development and psychometric characteristics of the Self-Description Questionnaire administered to sampled children, and chapter 7 presents the same information for the teacher indirect cognitive and social rating scale measures.

A national probability sample of about 22,000 children in about 800 public and 200 private schools was assessed at entry to kindergarten in fall 1998 (round 1). They were followed up in spring-kindergarten (round 2), fall- and spring-first grade (rounds 3 and 4, respectively), spring-third grade (round 5), and spring-fifth grade (round 6). The third round (fall-first grade) was a subsample of about 30 percent of the base-year kindergarten schools. The sixth round of data collection described in this report took place in spring 2004, when approximately 90 percent of the children were in fifth grade. The direct cognitive assessments were conducted in all six rounds of data collection, while the indirect cognitive and socioemotional measures were collected from teachers in rounds 1, 2, 4, 5, and 6 (fall- and spring-kindergarten, spring-first grade, spring-third grade, spring-fifth grade), and from parents in rounds 1, 2, and 4. In rounds 5 and 6, children completed a direct socioemotional measure. More details on the sample design and data collection methods used in the ECLS-K can be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Combined User's Manual for the ECLS-K Fifth-Grade Data Files and Electronic Codebooks* (NCES 2006–032) (Tourangeau et al. forthcoming).

Sample counts, completion rates, psychometric characteristics, and score statistics for the fifth-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. Statistics presented in this report may differ slightly from those in the data file users' manual. Tables in the users' manual are based on the panel sample, that is, children who participated in all six rounds of data collection, with national estimates computed using the longitudinal panel weight (C1_6SC0). The emphasis in this report is on the psychometric characteristics of the tests at each round, so all children participating in each round are included, and the corresponding cross-sectional weights, (C1CW0–

www.manaraa.com

C6CW0) are used for national estimates. Statistics that report characteristics of the tests rather than national estimates, such as reliabilities or floor and ceiling effects, are unweighted. Detailed information on the assessments used in the earlier rounds can be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99* (*ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062) (Pollack et al. 2005).

www.manaraa.com

2. DESIGN AND DEVELOPMENT OF THE ASSESSMENT INSTRUMENTS

The Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) assessment instruments were designed to measure children's academic and social development during the kindergarten through 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 sixth round of data collection, when most of the ECLS-K children were in fifth 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 ensure valid representation of domain content and to make efficient use of resources while minimizing burden on teachers and students.

The fifth grade direct cognitive assessments built on the structure established in the kindergarten through third-grade 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.

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

- **No English language screening**: In kindergarten and first grade, children who were identified as coming from a language minority background were administered an

www.manaraa.com

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 were 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 after first grade, the number of sampled children who might still lack English proficiency two and four years later, in third and fifth grades, was assumed to be so small that the language screening assessment would be unnecessary. Therefore, an English language screener was not administered after spring-first grade.

- **No parent questionnaire items on children's social behaviors:** Parents' ratings of children's behavior and social skills had been collected during kindergarten and first-grade rounds. These ratings were deleted from parent information collected in third and fifth grades 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 rounds of data collection.

- **Self-Description Questionnaire (SDQ):** In third and fifth grades, 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.

- **Changes in the content and format of the direct cognitive assessment instruments:** New reading, mathematics, and science assessment forms were developed for the fifth grade. A science assessment, begun in third grade, replaced the direct cognitive assessment of general knowledge that had been used in kindergarten and first grade. Assessment formats in fifth grade 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 and fifth grades.[1] Additional scores were defined that targeted fifth-grade skills. Details of these changes are described in sections 2.1 and 2.2.

- **Changes in the indirect cognitive assessment instruments:** Separate teacher ratings of science and social studies skills in third grade replaced the K-1 general knowledge ratings. In fifth grade, the social studies section was eliminated in order to reduce teacher burden.

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

www.manaraa.com

Another change in the longitudinal design of ECLS-K was the elimination of the second- and fourth-grade rounds of data collection due to budgetary constraints. The implications of this decision, and the steps taken to minimize its impact on longitudinal measurement, are discussed in sections 2.1.5 and 5.1 of the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062) (Pollack et al. 2005).

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 direct cognitive assessment), mathematics, and science were selected for the fifth-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 in reading and mathematics throughout the whole span of the study, and in science between third and fifth grades. The longitudinal design of the study required the development of vertical scales in each subject 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. Some compromises were necessary to reconcile the goal of using age-appropriate reading passages with the objective of limiting total test time to an average of 75 minutes in fifth grade. The time limitation precluded the use of assessment tasks such as extended reading materials or hands-on science experiments.

As noted earlier, the same reading, mathematics, and general knowledge assessment instruments had been used in all four kindergarten and first-grade rounds of data collection. Children were routed to different levels of difficulty within each assessment domain depending on their performance on

www.manaraa.com

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 and fifth grades could be expected to have advanced beyond the levels covered by the K-1 assessments, new sets of assessment instruments were developed for each round after those for the first grade. Some test items were retained from each round to the next to support development of a longitudinal score scale.

The K-1 general knowledge assessment, which included basic natural science concepts as well as concepts in social studies, was replaced by a direct cognitive science assessment administered in third and fifth grades. The science assessment is not comparable to the K-1 general knowledge assessment, so the longitudinal scale in science spans only the last two rounds of data collection. As a result, 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 fifth-grade assessment was similar to that of prior rounds, with some changes to accommodate the more advanced level of the questions. As in the earlier years, an assessor presented the questions to the child and entered responses into a computer for each individually administered assessment. Seven of the mathematics items asked the child to carry out a task, such as completing a graph or diagram, measuring an object, writing a decimal number, or solving a problem requiring a computation. These items were administered in workbook format. Each child received between one and six workbook items, depending on which second-stage form was selected. To accommodate the length of the reading materials used in the fifth-grade assessment, a separate booklet containing both the reading passages and questions was given to the student, with the questions also appearing on the easel handled by the assessor. Section 5.1.2, Evaluating Common Items, describes the procedures used to evaluate common-functioning of items across different assessment rounds.

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- and fifth-grade batteries were administered entirely in English.

www.manaraa.com

The types of scores reported for the fifth-grade direct cognitive assessments are similar to those for kindergarten through third grades, with some modifications for scores representing both broad-based measures and targeted skills. Assessment scores were recalibrated and rescaled for fifth 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 fifth grade, and in science for third to fifth grade. As a result, scores in the public-use files for the earlier rounds should not be compared with recalibrated/rescaled scores in the kindergarten through fifth-grade public-use file. Scores from the earlier rounds that are required for longitudinal measurement have been rescaled and appear in the kindergarten through fifth-grade file in a metric that makes comparisons possible. New targeted scores based on clusters of fifth-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,

www.manaraa.com

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 both individually administered and adaptive. Individual administration of assessments was retained in fifth grade, even though children would probably have been able to cope with paper-and-pencil test forms at this time. The success of the adaptive approach in earlier rounds in optimizing measurement characteristics for a diverse sample of children suggested its use in the later grades as well. A change to group administration was considered for third and fifth grades 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,

www.manaraa.com

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 and fifth-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 after first grade) than for reading or mathematics. Second, the effect found was so small that it was inconsequential. Refer to the *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, 2 years after the design phase, and continued 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

www.manaraa.com

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

www.manaraa.com

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 through 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. The items for the kindergarten and first grade are allocated according to the amount of time items were expected to take. However, the content items for the third and fifth grades are distributed by the percentage of items to match the NAEP frameworks.

2.1.2.1 Reading Test Specifications

The ECLS-K reading specifications were adapted from the 1992 and 1994 NAEP Reading Frameworks (National Assessment Governing Board [NAGB] 1994a). 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

www.manaraa.com

story types allow us to get information on critical skills as early as kindergarten. Third- and fifth-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 the 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. Nevertheless, 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).

www.manaraa.com

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

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 content strands are identical to the National Assessment of Educational Progress 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, spring 2002, and spring 2004.

www.manaraa.com

2.1.2.2 Mathematics Test Specifications

The mathematics test specifications shown in table 2-2 are primarily based on the Mathematics Framework for the 1996 NAEP (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 **r**epresenting 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. 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.

www.manaraa.com

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

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, spring 2002, and spring 2004.

www.manaraa.com

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 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 and fifth grades. No direct measurement of social studies knowledge was included in third and fifth grades, although teacher ratings of children's proficiency in social studies were collected in third (but not fifth) grade. 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 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

www.manaraa.com

weather and climate; events include volcanic eruptions, earthquakes, and storms. Space science in the 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 assessed relate to interdependence, adaptation, ecology, and health and the human body.

Table 2-3. Science longitudinal test specifications, in percent of test items, for third grade (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 content strands and target allocations. The allocation of items at each grade level follows the 1996 NAEP guidelines that specify that about half of the items within each of the science subdomains measure conceptual understanding and half measure scientific investigation. Detail may not sum to 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 and spring 2004.

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 results, and sets of questions were selected

www.manaraa.com

for a full-scale field test in spring 2002. Fourth-graders were included in the field test sample along with fifth-graders, in the event that a fourth-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 fifth-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 (scored right/wrong), 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, and of fourth and fifth-graders in 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 fourth and fifth grade was designed primarily to gather the

www.manaraa.com

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 was chosen from several instruments reviewed because the content it covered, the time it took to administer, and its available reliability and validity information best suited its use as a validation instrument for ECLS-K. The MBA subtests were administered according to standard procedures specified by the publisher.

Spring 2002 field test. Between 120 and 136 questions in each of the content areas (i.e., reading, mathematics and science) were field tested. Most of the field test items were taken from the 1999 pilot tests, with revisions incorporated as necessary. Four reading passages and the accompanying items, which had been administered to eighth-graders in the National Education Longitudinal Study of 1988 (NELS:88) were also field tested. These were selected to serve two purposes: to supply high-difficulty questions that were based on relatively short reading passages, and to facilitate linking to NELS:88 score scales if ECLS-K were to be extended beyond fifth grade. The items within each of the content areas were divided into two parallel sets of items, A and B, with separate workbooks accompanying the mathematics sets for 12 (Form A) or 10 (Form B) of the 60 mathematics items. Six booklets, each containing two subtests in different content areas, were created (table 2-4).

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 2002 field test

Field test form	Section 1	Section 2	MBA Validation
1: Red	Mathematics A (60)	Reading A (68)	(none)
2: Orange	Mathematics B (60)	Reading B (68)	(none)
3: Yellow	Science A (60)	Mathematics A (60)	Mathematics MBA (29,50)
4: Green	Science B (62)	Mathematics B (60)	Mathematics MBA (29,50)
5: Blue	Reading A (68)	Science A (60)	Reading MBA (28,22,23)
6: Purple	Reading B (68)	Science B (62)	Reading MBA (28,22,23)

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 2002 field test.

www.manaraa.com

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.

At the end of four of the six booklets, an MBA reading or mathematics test was administered for validation purposes. In each case, the booklet in which a set of science items was paired with reading or mathematics was selected for administration of the corresponding MBA test. The first two booklets, red and orange, did not include an MBA validation section. Each of these two booklets, which paired a reading with a mathematics section, was already quite long and challenging for the children because of the time required for reading passages and mathematics computations. The science sections, with only short-answer questions, were faster to administer and left time for an MBA test at the end.

About 1,800 children, approximately evenly divided between fourth- and fifth-graders, participated in the field test of cognitive items in spring 2002. 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 fourth-graders and half from fifth-graders.

Early in the field test period, it became clear that the test forms were too long for some children to complete in a reasonable period of time. Modifications were implemented to minimize burden on the children, while ensuring that sufficient data would be collected for the purposes of selecting test items and validating the item pool. Two reading passages were deleted from the test forms for the remainder of the field test period, one because it was excessively long, and the other because it was too easy. One mathematics item, which had proven to be flawed, was also deleted. Discontinue rules were defined so that if an excessive amount of time had elapsed at certain checkpoints within the booklets, the assessor would then skip to a later item set within the same section, to the next section in the booklet, or to the MBA validation section that followed the field test item sets.

www.manaraa.com

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 cooperation. Comprehensive reports from the assessors who administered the field tests complemented the analysis of item response data, and played an important part in the design of the fifth-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 fourth and fifth grade participants combined, as well as for each grade separately. Item discrimination, that is, the extent to which each item is consistent with the overall set of items, was measured by *r*-biserials, which are correlations of total score with item score (right/wrong) for each item. Distractor analysis consisted of evaluating statistics on the percentage of children choosing each response option for multiple-choice items, and the average total test score for those choosing each option. This information provided a basis for identifying items in need of revision, for example, questions that might have more than one potentially correct answer, incorrect response options chosen by children scoring higher, on average, than those choosing the intended correct option, 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 subsequent test questions, can be an indication that a test item is confusing or otherwise problematic for children. Classical item statistics also included the alpha coefficient, a measure of reliability, for each set of field test items.

IRT parameters (Lord 1980) were estimated for all cognitive items in the field test, using the PARSCALE computer program (see section 3.2.2 for details) for the purpose of item selection only. (Parameters were re-estimated later using national sample data.) The IRT parameters were based on the three-parameter model with a parameter for guessing, a parameter for difficulty, and a slope (discrimination) parameter. The IRT slope, or "a" parameter, complements the information provided by the *r*-biserial but relates item discrimination to overall performance at a particular ability level rather than for the whole range of ability. The "b" parameter provides a measure of difficulty that is less susceptible

www.manaraa.com

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. ICCs were inspected for indications of lack of fit. Graphs containing the ICCs also included markers showing percent correct, separately for fourth and fifth graders, at intervals spaced along the ability range. This permitted evaluation of overall fit as well as displaying possible differences in functioning for the two grades. A relatively small percentage of items exhibited overall lack of fit and were removed from consideration for the fifth-grade battery. Examination of ICCs for the poorer fitting items, along with the distractor 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 fifth-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 fifth-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 fifth-grade forms.

The question of whether the absence of a fourth-grade round of data collection might result in a gap in ability levels that might seriously impact the measurement of gain was addressed. Examination of the field test results showed a considerable overlap in ability distributions between third- and fifth-graders. As a result, no fourth-grade "bridge" data collection, analogous to the second-grade sample that had been assessed to bridge the first-to-third grade gap, was necessary.

Cognitive test items were checked for DIF for males compared with females. There were too few Hispanic and Asian children in the field test sample for DIF analyses to be carried out for these groups. Sample sizes of Black students were sufficiently large for Black/White DIF to be evaluated for only about half of the field test items. 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.

www.manaraa.com

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

A finding of differential functioning, however, does not automatically mean that a test item is inappropriate. It simply means that the item is differentially easier or more difficult for some subgroup (focal group) when compared with a reference group. A judgment that an item is inappropriate requires not only the statistical measure of DIF for one or more subgroups, but also a determination that the difference in performance is *irrelevant to the construct being measured*. In other words, different population subgroups may have differential exposure or skill in solving test items relating to a topic included in the test specifications. If so, the finding of differential performance may be an important and valid measure of the targeted skill, and should be included in the assessment (see section 3.4; also Holland and Thayer 1986). Items that demonstrate differential functioning favoring the reference group were reviewed for inappropriate content by a standing committee on test fairness at Educational Testing Service (ETS), consisting of both majority and minority group members. Items that were judged to have content or presentation that might be problematic for a particular focal group in ways that are not relevant to the construct being measured were dropped from the item pool. For example, a mathematics item requiring students to mark the location of an ordered pair on a grid turned out to be differentially more difficult for Black compared with White students, while a science question based on relative weights of three blocks was differentially more difficult for females compared with males. 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 a finding of "C" level DIF for four reading, three mathematics, and two science items. The mathematics and science "C" DIF items, and two of the four reading items, were deleted from consideration for the fifth-grade assessments. The remaining two reading "C" DIF items were retained: one had been previously reviewed in the third-grade assessment and found to be acceptable; for the other, statistics suggested that the DIF finding might be merely an artifact of small sample size.

www.manaraa.com

Correlations between total reading and mathematics scores on the MBA construct validation instrument and the corresponding field test reading and mathematics IRT ability estimates (thetas) were computed. The correlations (.73 for reading, .80 for mathematics) were somewhat lower than the MBA correlations for the corresponding subjects in the second- and third-grade field test (.83 and .84, respectively). Two factors contributed to the lower correlations between reading comprehension measures:

- **Differences in content between MBA and ECLS-K tests increased from the 2000 to the 2002 item pools:** While the same MBA tests were used in both field tests, the ECLS-K item pools were different. The MBA tests place a great deal of emphasis on basic skills, while the ECLS-K pools moved toward increasing emphasis on reading comprehension and mathematics problem solving. MBA reading sections contain 28 "identification" (decoding items), 22 vocabulary items (opposites), and 23 comprehension items. About two-thirds of the ECLS-K grade 4 to 5 reading field test items were comprehension questions based on reading passages. Similarly, the MBA mathematics test contained 29 calculation questions (most children discontinued the section after item 18 or 19) and 50 "reasoning and concepts" items, of which most children answer about 20 because the easiest and hardest items were not administered. The ECLS-K field test grade 4 to 5 mathematics item pool contained very few pure calculation questions and a large majority of word problems.

- **Reduced MBA score variances:** The standard deviations of the MBA scores were somewhat lower for the grade 4 to 5 field test than for grade 2 to 3. All other things being equal, reducing variance results in lower correlations. (The ECLS-K ability estimates did not have a lower variance for grade 4 to 5, in fact, variances were increased.) Early in the field test, a proposal to save field test time by deleting the MBA section for children who were taking a long time to complete their ECLS-K item sections was rejected because of the need to preserve the variability of the MBA sample. There is evidence that these lower-scoring children *did* receive the MBA: the mean ability estimate for the half-sample of MBA takers was similar to the mean for the whole field test sample.

Although the correlation coefficients were lower than those found for the grade 2 to 3 field test, the correlations of .73 for reading total and .80 for mathematics total are sufficiently high to support the purpose of validating the ECLS-K item pools.

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.

www.manaraa.com

Start and stop times were recorded for each field test section as a whole, and for reading sections, separately for each reading passage as well. As noted earlier, the field tests proved to be too long for the time allotted and for children's ability to function effectively. The deletions and discontinue rules instituted to shorten the test meant that timing data were relevant only for test sections that appeared in the first position in the booklet, before stop rules impacted the collection of second-section data for some children. Item nonresponse rates for sections given first were low: relatively few children either omitted items while answering subsequent questions or failed to reach the end of the test sections.

While the timings for mathematics sections were somewhat longer than expected, the reading sections were primarily responsible for the excessive field test times, for several reasons. Unlike the mathematics and science questions, which were independent short answers or computations, most of the reading questions required the additional investment of time to complete the reading passages on which the questions were based. Analysis of timing data showed that roughly half of the total time was used in reading, the other half in answering the questions. Another factor responsible for the excessive time required for the reading sections was the length of the reading passages themselves. In an effort to make the assessments reflect tasks typical of the fifth-grade curriculum, passages of several pages in length were included in the field test. Timing results made it clear that given the time constraints of the assessment, the requirement of curriculum relevance must be satisfied by the difficulty of the reading materials and questions rather than their length. In interpreting results for the longer and more difficult reading passages, it is also important to take into account the difference in method of assigning forms in the field test (random) compared with the fifth-grade operational test (form selection based on ability demonstrated in the routing section). Passages that were too long and difficult for the randomly-assigned field test participants may be suitable for the reading form designed to be administered to the highest achieving fifth-graders.

Timings for the mathematics and science sections were recorded for the section as a whole, not for individual items or groups of items. On average, the mathematics sections took under 40 minutes to complete, and science sections under 32 minutes, or about two-thirds of a minute per item for mathematics and one-half minute per item for science. This suggests that 35 to 40 questions per child could be administered in the national test within the target time of 30 minutes for mathematics and 20 minutes for science.

Field test assessors participated in debriefing sessions following the spring 2002 field test administration. They provided information on the children's reactions to test questions as well as

www.manaraa.com

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 on questions and response options that were confusing, ambiguous, or incomplete were taken into consideration in selecting and revising items for the proposed fifth-grade reading forms. Comments related to performance (such as reports that children found an item too difficult) were, in general, corroborated by analysis of the field test data, although some reading passages that assessors reported that they or the children did not like did have satisfactory statistics. This was particularly true of the long reading passages, which the less able readers clearly found too challenging. The most important point made by the examiners was the need to keep testing time short enough so that children would not get tired and frustrated.

The booklet design described earlier, with each test form appearing both early and late in a testing session, was designed to permit analysis of order effects. However, the discontinue rules implemented to shorten the assessment resulted in many children—primarily the lower achievers—failing to complete the second section of the field test booklet. This made comparison of statistics for sets of items given early in the testing session with the same items given in a later position impossible. Assessors' reports clearly indicated that fatigue due to test length was a factor in performance. In the second- and third-grade field tests, when excessive test length was not an issue, neither a practice effect (better performance toward the end of the test) nor a fatigue effect (a drop in performance) was found.

2.1.4 Fifth-Grade Test Forms

The fifth-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 fifth-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 fourth- and fifth-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.

www.manaraa.com

2.1.4.1 Item Quality and Reliability

To contribute useful information about children's skill levels, test items selected for the final forms should ideally have high *r*-biserials (.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 third-grade assessments, or links to a selected reading passage. In IRT, the measurement precision for individual examinees is improved by administering the maximum number of items possible in the time available, and including items that function appropriately and measure the same construct. Items found to have DIF for population subgroups were deleted from the item pool except as noted earlier.

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 fifth-graders was used to determine item difficulty objectives such that the middle-difficulty form would be suitable for approximately the middle half of fifth-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 fifth-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 fifth-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.

www.manaraa.com

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 fifth-grade assessment forms were designed to have enough easy items that distinctions could be made at the low end of the ability range, and enough hard items to accurately measure the most skilled students. To avoid floor and ceiling effects, each assessment included a few items in the high second-stage form that almost all children would get wrong, and a few in the low second-stage form that almost all children would get right, so that accurate measurement of the extremes of ability could be accomplished.

Each of the second-stage test forms contained some items with difficulty levels that extended beyond the target ability range, at both the high and low end. This design feature served two purposes. First, it provided some of the overlapping items required to put all of the test forms on a common scale (in addition to routing items taken by all children). Second, it improved measurement properties for children whose achievement level was very near a routing cut point. There was the possibility that guessing and/or careless mistakes on the routing test could result in children at the margin receiving a second-stage test form that was too easy or too hard. For example, a child whose ability level was 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 was 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- and fifth-grade assessments used the same overlapping two-stage design but with more advanced sets of items. Putting K-1, third-, and fifth-grade scores on a common scale required

www.manaraa.com

common items shared between subsequent assessments. Items from the K-1 assessments (22 in reading, and 14 in mathematics) provided the necessary link between K-1 and third grade, with a small "bridge" sample of second-graders augmenting the gap in ability levels between first and third grade. Overlapping ability distributions for third- and fifth-grade made a fourth-grade bridge sample unnecessary. Fifth-grade items shared with the third grade assessment (59 common items in reading, 31 in mathematics, and 27 in science) supported the extension of the K-1-3 longitudinal scale through fifth grade.

2.1.4.5 Curriculum Relevance

Both fourth- and fifth-graders participated in the 2002 field test of cognitive items. Although there was no fourth-grade round of data collection, the fourth-grade field test data did play a role in the design of the test forms for the fifth grade. Analysis of field test data was carried out for both grades combined, as well as separately for fourth grade and fifth grade. In selecting items for the fifth-grade test forms, preference was given to items that showed the largest differences in percent correct between the fourth- and fifth-graders in the field test sample. Although the fourth- and fifth-graders in the field test were different children, not longitudinal measurements of the same children, items with the largest fourth-grade to fifth-grade differences in percent correct could be assumed to be strongly related to fifth 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 fourth-grade and fifth-grade field test participants, suggesting that their content was not emphasized in fifth-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 the earlier rounds, avoiding floor and ceiling effects, and maintaining item quality. This was especially true for the reading assessment in which several questions based on each reading passage placed an additional constraint on the selection of items to match content strands. Reading items were not selected individually but in sets of four to eight 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 high quality items as possible based on the passage, even if that resulted in overrepresentation of

www.manaraa.com

a content strand. Conversely, a shortfall in a content strand could result if the available items in the strand were linked to a reading passage that had too few other useful items to justify its selection.

Table 2-5. Reading fifth-grade framework targets and percent of assessment items: School year 2003-04

Percent of assessment items	Total	Basic skills	Vocabulary	Initial understanding	Developing interpretation	Personal reflection	Critical stance
Target	100	10	10	15	30	15	20
Actual	100	17	11	23	26	5	18

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

Table 2-6. Mathematics fifth-grade framework targets and percent of assessment items: School year 2003–04

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	42	23	12	8	15

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

Table 2-7. Science fifth-grade framework targets and percent of assessment items: School year 2003–04

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

NOTE: Detail may not sum to totals due to rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

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 the earlier rounds and avoiding floor and ceiling 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, and 10 percent in fifth

www.manaraa.com

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. While the decoding items did provide a valuable link between third and fifth grades, the presence of many common items from the reading passages shared between third and fifth grades made this issue less critical. A more important reason for selecting additional Basic Skills items was the need to fill gaps in the distribution of item difficulties. Since these items were not tied to reading passages, they could be selected individually at points where items of a particular difficulty were needed. In fact, the impact of the overrepresentation was minimal for two reasons. First, the 16 Basic Skills decoding items were administered in sets of four, in ascending order of difficulty, and the harder sets were skipped if a performance criterion on an easier set was not met. The last set of four items was administered to only 13 percent of the children. Second, an adjustment was made when scores were calculated. All of the Basic Skills items were utilized in estimating ability levels, but four were deleted from computation of the final scale scores to align the composition of the scores more closely with the framework.

Initial Understanding items were overrepresented in comparison to framework targets, while Personal Reflection items were underrepresented. Nearly half of the Initial Understanding items were selected for the lowest fifth-grade form and served the purpose of linking to the earlier rounds. The remaining Initial Understanding items were retained because they accompanied a selected reading passage. The shortage of Personal Reflection items, as for the third-grade assessment, was due to relatively poor psychometric performance for items in this category.

Item selections for the mathematics and science assessments closely matched framework target percentages, in large part because the constraint of selecting items in groups was not present. Enough high quality science items were available for selection in each of the content strands to match frameworks exactly, with only minor deviations from targets in mathematics.

The deviations from framework targets probably have relatively little impact on the measurement of the domain of interest because there is some ambiguity in the classification of items. Many if not most of the 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

www.manaraa.com

numbers. Therefore, the designation of a single strand category for each item was somewhat arbitrary. It is unlikely that the necessary compromises in selecting items would have a serious negative impact on measurement of the intended construct.

2.1.4.7 Practical Issues

The 75-minute time allocation for the fifth-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 fifth-grade test forms is shown in table 2-8.

Table 2-8. Number of items in fifth-grade test forms and routing test cut scores, by domain: School year 2003–04

Description	Reading	Mathematics	Science
Number of items per form			
Routing test	26 (25 scored)	18	21
Low second-stage form	24	18	15
Middle second-stage form	25	18	17
High second-stage form	31	19	14
Total number of items			
Fifth-grade pool	94	60	57
K-1-3 Scale (Science third grade only)	154	123	62
Overlap between K-1-3 and fifth grade	59	31	27
Items in longitudinal scale (K-1-3 -5)	186	153	92
Routing test cut scores			
Route to low second-stage form	0–8	0–8	0–8
Route to middle second-stage form	9–16	9–13	9–14
Route to high second-stage form	17–26	14-18	15–21

NOTE: The number of items in each fifth-grade pool is less than the sum of the items in the test forms because there is some overlap of items across forms. Four fifth-grade reading items were calibrated but deleted from the final score scale to align the scale with the framework, and one was deleted from scoring because of differential item functioning (DIF) in the fifth-grade sample. Two reading items that had not been scored in third grade because they proved to be too difficult to provide useful information for third-graders performed satisfactorily when fifth-grade responses were added to the analysis. These two items, present but not scored in third grade, were added to the longitudinal scale. Similarly, one mathematics item that had unsatisfactory statistics in third grade was added to the longitudinal scale based on the combined third and fifth-grade data. See chapters 4 and 5 for details.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

www.manaraa.com

Routing test cut points were determined empirically based on field test IRT ability estimates and item parameters. Using the ability estimates for field-tested fifth-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. Cross-tabulations of the simulated routing scores against each second-stage score were examined, and routing cut points were selected such that ceiling and floor effects would be minimized. 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 fifth-grade longitudinal sample. The success of the two-stage test design in achieving its goals is discussed there as well.

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*-biserials) corroborated their reports.

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

2.2 Indirect Measures: Teacher Ratings

Teachers of ECLS-K children in previous data collection cycles received two questionnaires (A and B) asking about their background, training, and classroom practices. They also received a third

www.manaraa.com

questionnaire (C) that asked the teacher to rate each ECLS-K child on sets of academic and behavioral measures. By fifth grade many schools no longer have self-contained classrooms and students may be taught by two or more teachers. Therefore, the child's reading teacher completed the questions about the child's language and literacy and social development, as well as providing information about the child's classroom experiences in language and literacy. The questions on classroom experiences addressed both the classroom and student peer characteristics, as well as the instructional and curricular aspects of the classroom. Separate questionnaires pertaining to mathematics performance and instruction and science performance and instruction were provided to the child's mathematics teacher and science teacher respectively. To reduce the cost of data collection, children were randomly assigned to two of the three questionnaires. All children were rated on the language and literacy measure. Therefore, children were rated by teachers on language and literacy, social skills, and mathematics OR language and literacy, social skills, and science. All children were rated by the reading teacher on the social-emotional scale. The following two sections describe the indirect assessments of children's academic performance and social-emotional development that the teachers completed.

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 social studies domain was not included in the fifth-grade data collection. The ARS was designed both to overlap and to augment the information gathered through the direct cognitive assessment battery. Although three 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. 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 used 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.

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 multi-digit problems

www.manaraa.com

with remainders in the quotient") 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 (e.g., "Demonstrates money management skills, for example, computes savings on a 20 percent off sale, balances a classroom savings account, or determines profit earned on candy bar sales"). 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 scope and sequence in published mathematics curricula and the literature on the importance of skills at different grade levels 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.

www.manaraa.com

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). 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), Combined User's Manual for the ECLS-K Fifth-Grade Data Files and Electronic Codebooks* (NCES 2006–032) (Tourangeau et al. forthcoming).

Exhibit 2-1. Academic Rating Scale response scale, fifth grade: School year 2003–04

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 2004.

Teachers from both public and private schools and from different regions of the country and content experts familiar with the elementary 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 set rather high in order to capture the range of abilities represented in fifth grade and to avoid a serious ceiling problem. The items were field tested in the spring of 2002, at the same time as the field test of the direct cognitive assessments. One fourth-grade teacher and one fifth-grade teacher in each of the 49 participating field test schools were asked to participate in the field test by completing five child rating forms: one for the highest achieving child in their class, one for the lowest achieving child in their class, and three for children with average achievement, regardless of whether these particular children were participating in the direct assessment. Participating schools were alternatively assigned "blue" and "white" designations to ensure equitable distribution of the two forms of teacher ratings. The covers of the teacher questionnaire were blue or white. At "blue" schools, teachers were asked to rate these children

www.manaraa.com

using the ARS for the children's current grade level and the grade level below. At "white" schools, teachers were asked to rate children using the ARS for the children's current grade level and the grade level above. A total of 545 teachers, 277 fourth-grade and 268 fifth-grade teachers, completed field test forms. 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. The third grade version of the SRS was administered in fifth grade. 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 Combined User's Manual for the Fifth-Grade Data Files and Electronic Codebooks* (NCES 2006–032) and the *ECLS-K Psychometric Report for the Third Grade* (NCES 2005–062).

Exhibit 2-2. Social Rating Scale response scale, fifth grade: School year 2003–04

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 2004.

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

www.manaraa.com

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- or fifth-grade parent interview.

2.3 Self-Description Questionnaire

In the third-grade data collection and again in the fifth grade, students rated their own academic competence and social skills. The 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 and spring 2002 field tests: Reading, Mathematics, All School Subjects, and Peer. The students responded to the SDQ-I questions prior to the administration of the cognitive assessment. The response scale as well as several of the items were adapted for use, with permission, in the main study and administered in the third- and fifth-grade data collection periods.

The original SDQ-I has some negatively worded items that were not scored, but were included in the instrument 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 field tested in spring 2000 and spring 2002.

After analyzing different combinations of responses, it was found that a three- to four-point response scale worked best. A four-point scale offered the opportunity to get as much variance as possible within the ability of third- and fifth-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,"

www.manaraa.com

"a little bit true," "mostly true," or "very true." This also reduced the cognitive demand for the students. The same scale was used in third and fifth grades.

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 and sixth rounds 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. (6 items)
- **SDQ Sad/Lonely/Anxious** scale includes items about internalizing problem behaviors such as feeling "sad a lot of the time," feeling lonely, feeling frustrated, feeling ashamed of mistakes, and worrying about school and friendships. (8 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."

www.manaraa.com

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. The third-grade instrument was used in the fifth grade. However, in the fifth-grade the item asking about frustration loaded more heavily on the Sad/Lonely/Anxious scale, while in the third grade it had been more closely related to Anger/Distractibility.

www.manaraa.com

3. ANALYSIS METHODOLOGY

This chapter describes the procedures used in processing the ECLS-K fifth-grade assessment data and producing scores for analysis and for inclusion in user files. Quality control steps are described in section 3.1, followed by an explanation of the methodology used to carry out specialized procedures for psychometric analysis. A three-parameter item response theory (IRT) model was used to put scores obtained on different assessment forms on the same scale for the purpose of comparisons within and across assessment years. The Rating Scale model (Wright and Masters 1982), 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. The development of longitudinal score scales is described in chapter 5.

3.1 Quality Control Procedures

Procedures employed to ensure accuracy in the collection of the cognitive test item data are described in section 4.6 of the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Fifth-Grade Methodology Report* (NCES 2006–037) (Tourangeau et al. forthcoming). In the subsequent steps of converting the resulting raw item response data to final scores, procedures were checked to ensure the accuracy and validity of the results. A series of steps were carried out, from converting raw examinee item responses into scores for individual items, to evaluating item functioning using both classical item analysis and IRT methods, to assembling item data into meaningful and interpretable scores. Throughout the process, attention was given both to checking that steps were carried out correctly, and to verifying that results accurately represented the constructs they were designed to measure.

Frequency distributions of raw examinee item responses were produced for each test item to serve as a baseline for confirming the accuracy of later processing steps. Each distribution was compared with the text of the corresponding question in the assessment easel, and with the instructions the assessor used in recording responses, to confirm that responses were coded as expected. For example, for a four-option multiple choice question, the data file would be expected to contain response codes of 1, 2, 3, and 4, while 1 (correct) or 2 (incorrect) was to have been recorded by the assessor for open-ended questions. Missing data codes (8 = refused, 9 = "I don't know") were also counted for each item.

www.manaraa.com

Within each subject area, children who had not responded to enough test items to receive a score were identified. "Too few items" was defined as answering fewer than 10 questions in the routing and second stage forms combined. For the purpose of identifying unscoreable cases, codes for "I don't know" were *not* treated as valid responses. Only items actually attempted by the child were counted toward the scoreability threshold. Before being deleted from further analysis, each "too few items" data record was reviewed visually to verify that not enough valid item responses were present.

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

Summary statistics for each item include P+ (percent correct) and *r*-biserial (the correlation of item score with total test score, adjusted for the item score being dichotomous). These statistics were reviewed to verify that an unambiguous correct answer key was used for each item, meaning not only that the *intended* right answer was tagged, but that the tagged answer was in fact functioning as an unambiguous right answer. Evidence for the validity of the answer key comes from two sources: the mean average section score for test takers choosing the correct response should be higher than that of the groups choosing incorrect responses; and the *r*-biserial should be positive, ideally at least .30 or higher. If these conditions are not satisfied, one of two error conditions could be responsible. An incorrect answer key could have inadvertently been applied or the item may be flawed; that is, the intended correct answer may not really be correct, or there may be two or more equally correct response options. Because all of the fifth-grade items had been field tested and response options evaluated and corrected, if necessary, no flawed items were found.

Items within each test section had been arranged in ascending order of anticipated difficulty. A review of the item P+s would identify any serious deviation from this expectation, which could indicate

www.manaraa.com

anomalies in the administration or scoring of items. Similarly, unexpectedly large omit or not-reached counts for an item or items could call into question whether routing steps or discontinue rules were applied correctly. No such indicators of data or administration errors were detected in reviewing item analysis tables.

Summary statistics from the item analysis include the number of items and number of test takers analyzed for each section, the highest and lowest scores encountered on the section, a measure of reliability (alpha coefficient), and a frequency distribution of the number right for the section. Reliabilities were reviewed to confirm that they were consistent with expectations: typically about .80 or above for routing sections and sections with more items, and lower than that for sections with relatively few items, and for second-stage forms, for which the restricted variance in overall ability (relative to the whole sample) would be expected to result in lower alpha coefficients. The reliabilities for all test sections were consistent with these expectations. Item and sample counts, and score ranges, were checked for consistency with known values.

Frequency distributions of routing test scores were compared with the distributions for each second stage form to confirm that the routing had been carried out at the correct cut points, i.e., that the number of observations for each second stage form matched the number in the corresponding score range of the routing test. In a small number of cases (one mathematics test, and five science tests) children had answered enough items in the routing test (at least 10) to be considered scoreable, but no items in a second stage form. Data records were reviewed visually to confirm that the discrepant counts (e.g., number routed to the low form vs. the number who answered one or more items on the low form) reflected what was actually in the raw data files.

Frequency distributions of total number correct (routing plus second stage combined) were examined separately for each form combination (i.e., routing+low form, routing+middle form, routing+high form) to look for possible floor and ceiling effects. While this is not a quality control issue in the sense of verifying the accuracy of the scoring procedures, it does have implications for interpretation and analysis of the resulting scores. A floor effect occurs when the test is too difficult overall for some test takers, and the score distribution contains a substantial number of children scoring at the chance, or guessing, level. Conversely, a test with a ceiling effect is too easy for some children, and a substantial number are able to answer all, or nearly all, of the items correctly. Only one set of tests, the science assessment for the children routed to the low form, had a floor effect, with about 5 percent of children scoring at the chance level (see section 4.5.1). This should have relatively little impact on

www.manaraa.com

analysis of scores because the IRT score calibration tends to shrink extreme scores to reflect the ability distributions for each round (see section 3.2.2 for a discussion of the effect of the Bayesian approach on chance and perfect scores). No evidence of a ceiling effect was found for any of the fifth-grade tests.

The next step in processing the raw item responses was preparing scored item files for input to the IRT calibration procedures, that is, replacing raw response option codes (e.g., 1, 2, 3, 4) with standard codes for correct, incorrect, omitted, and not reached items (1, 0, 2, and 3, respectively). Omitted items were defined as unanswered items that were followed by a response to at least one subsequent item, while unanswered items coded as "not reached" had no subsequent items answered. The quality control procedure for confirming that this was done correctly consisted of printing, for a spaced sample of every 1000th case, the raw and scored data record, along with the answer keys, and hand-checking the conversions. In some cases, additional records were needed, so that all variations found in the raw data file could be checked. For example, if the spaced sample of quality control records happened to have only cases that were routed to the low and middle second stage forms, additional records were obtained so that high form score conversions could be verified as well. Producing the scored item files entailed reorganizing the order of test items, because some items appeared in more than one second stage form. In order to strengthen the linkage of each set of forms to the same scale, the scores for these common items needed to be relocated from their original separate locations to a single common location. An item map was developed to direct the reordering of the common items. Scores that were simple sums of number correct on a specified set of items (reading and science cluster scores, reading and mathematics proficiency level scores: see sections 4.1.3 and 4.1.4 for definitions) were computed at this time, checked for the same spaced sample, and inserted into the scored item records. Although number-right proficiency scores do not appear in the user files because the sets of items were not taken by all test takers, the number-right counts for the proficiency levels were needed as input to the IRT calibration step. The fifth-grade scored item files were then combined with the scored item files from kindergarten through third grade. Like the test items shared in common across test forms within fifth grade, items shared in common across rounds were positioned together for IRT calibration, and again, frequency counts were checked to confirm the accuracy of the files.

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

www.manaraa.com

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

PARSCALE produces counts, for each test item, of the number of responses, number of omits, number right, and number wrong found in the input scored data file. Percent correct for each item is also computed. These counts and percents were checked, item by item, against the statistics generated from the scored, reordered data file to confirm that the correct input file was used and that the information it contained was interpreted correctly.

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

The graphs generated in conjunction with PARSCALE are a visual representation of the fit of the IRT model to the data. The modeled IRT parameters for each item define the shape and location of a logistic function for the item, which is plotted on a graph. Percentages of observed correct responses for grouped points across the range of estimated ability levels are superimposed on the same graph. The closeness of fit of the data to the logistic function can be interpreted as confirming the appropriateness of the IRT model for scoring the tests. More detail on the IRT model is presented in section 3.2, and a full description of the use and evaluation of the IRT procedures in developing the longitudinal scale appears in chapter 5.

The final steps in producing the IRT-based scores consisted of aggregating probabilities of correct responses across the whole item pool in each subject for the scale scores, and obtaining weighted

www.manaraa.com

means of ability estimates for standardized scores that represented population estimates at each round. These were checked by printing a spaced sample of every 1000th data case, including item and ability parameter estimates, and hand-checking computations. As a final checking step, means and standard deviations of the final score record were obtained, and found to be consistent with expectations. For the scale scores, that would be scale score means that increased from round to round, with ranges that were consistent with the number of items in the pool for each subject. The standardized scores could be explicitly checked, since by definition their weighted mean should equal 50.0 and standard deviation 10.0 within each round.

3.2 Overview: The Three-Parameter Model

Measuring the extent of cognitive gains at both the group and individual level requires that the various kindergarten through fifth-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.

3.2.1 Overview of Item Response Theory

The underlying assumption of IRT is that a test taker's probability of answering an item correctly is a function of his or her ability level for the construct being measured and of one or more characteristics of the test item itself. The three-parameter IRT logistic model uses the pattern of right, wrong, and omitted responses to the items administered in a test form and the difficulty, discrimination

www.manaraa.com

power, and probability of guessing correctly, given the lowest level of ability, of each item, to place each test taker at a particular point, θ (theta), on a continuous ability scale. Figure 3-1 is an example of a graph of the logistic function for a hypothetical test item. The horizontal axis represents the ability scale, theta. Points along the vertical axis represent the probabilities of answering an item correctly given the level of ability (θ).The shape of the curve is given by the following equation describing the probability of a correct answer on item *i* as

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

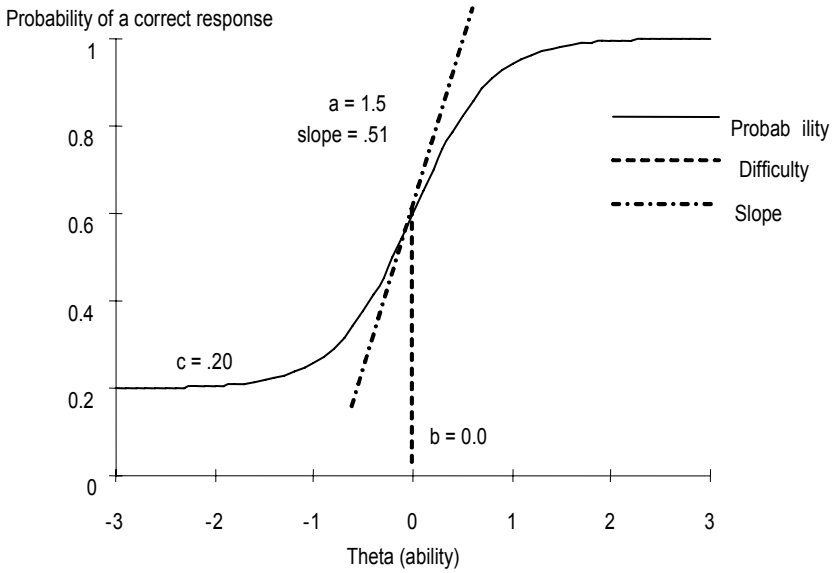
where θ = ability of the test taker;
a_i = discrimination of item i, or how well changes in ability level predict changes in the probability of answering the item correctly, at a particular point;
b_i = difficulty of item i; and
c_i = "guessability" of item i, that is, the probability that a very low-ability test taker will answer item i correctly.

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

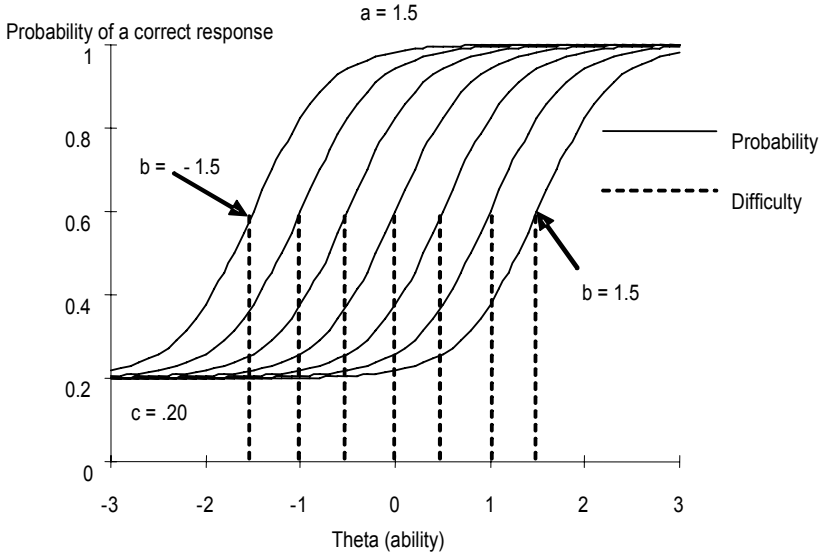
www.manaraa.com

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.

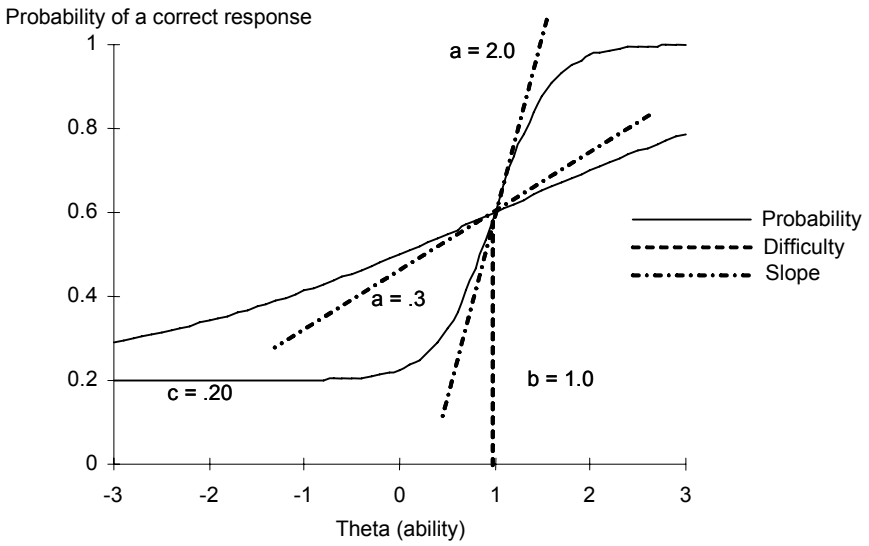
www.manaraa.com

The discrimination parameter, "a," has perhaps the least intuitive interpretation of the three IRT parameters. It is proportional to the slope of the logistic function at the point of inflection. Items with a very steep slope are said to discriminate well. In other words, they do a good job of discriminating, or separating, people whose ability level is below the calibrated difficulty of the item (who are much less likely to get it right) from those of ability higher than the item "b," who are much more likely to answer correctly. By contrast, an item with a relatively flat slope is of little use in determining whether a person's correct placement along the continuum of ability is above or below the difficulty of the item. This idea is illustrated by figure 3-3, representing the logistic functions for two test items having the same difficulty and guessing parameters but different discrimination. The test item with the steeper slope (a = 2.0) provides useful information with respect to whether a particular test taker's ability level is above or below the difficulty level, 1.0, of the item: if the answer to this item was incorrect, the person very likely has an ability below 1.0; if the answer 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" parameters, or item difficulties for the items, should span the range of abilities being measured. Item difficulties should be concentrated in the range of abilities that contains most of the test takers. Test items provide the most information when their difficulty is close to the ability level of the examinees. Items that are too easy or too difficult for most of the test takers are of little use in discriminating among them. Ideally the "c" parameters (the probability of a low ability person guessing correctly) tend to be about .25 or less for four-choice items, but they may vary with difficulty, and of course, the number of options. Open-ended items typically have a "c" parameter that is close to 0. In general, the ECLS-K item parameters met these standards.

www.manaraa.com

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



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

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

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

www.manaraa.com

3.2.2 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 estimates "a," "b," and "c" parameters for each test item, iterating until convergence when a specified level of accuracy is reached. Comparison of the IRT-estimated probability of a correct response with the actual proportion of correct answers to a test item for examinees grouped by ability provides a means of evaluating the appropriateness of the model for the set of test data for which it is being used. A close match between the IRT-estimated probabilities and the empirical probabilities means that the theoretical model accurately represents the empirical data.

As indicated earlier, a longitudinal growth study by its very nature consists of subpopulations defined by differing ability levels. That is, after all the kindergarten, first-grade, third-grade, and fifth-grade assessments had been completed (six rounds, counting fall and spring administrations in K-1) there are six 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 seventh subpopulation.

When the first round of kindergarten data was collected in fall 1998, relatively few children were routed to the middle-level second-stage forms and even fewer to the high level forms. Thus, there were not enough 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. The same is true for the most difficult first-grade items that were repeated in third grade, and for third-grade items repeated in fifth grade. 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, and now, with the addition of fifth-grade items to the scales,

www.manaraa.com

all previous rounds were re-estimated. The use of all data points over time is desirable because it can provide updated estimates of both the item and latent ability parameters throughout the entire ability distribution on a vertical scale. This procedure was used in the vertical scaling that was carried out for National Education Longitudinal Study (NELS:88) (Rock et al. 1995) and for High School and Beyond (Rock et al. 1985; Rock and Pollack 1987).

A strength of the PARSCALE and other Bayesian approaches to IRT is that they can incorporate prior information about the ability distribution (i.e., the round of data collection from which an observation is taken) in the ability estimates. This is particularly crucial for measuring change in longitudinal studies. It provides an acceptable way of coping with perfect and chance scores (i.e., correct answers to all items administered, or scores at the guessing level or below). 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. Conversely, individuals scoring at or below the chance level at two time periods may have gained skills that are below the level assessed by the test items. Pooling all available information, that is, pooling all item responses for all people at all time points, and re-calibrating 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 and chance scores, which in turn allows for the potential of some gains even in the upper and lower tails of the distribution. Each of the rounds of data collection in kindergarten through fifth 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.

www.manaraa.com

A multiple group version of the PARSCALE computer program (Muraki and Bock 1991) that was developed for the National Assessment of Educational Progress (NAEP) allows for both group ability priors and item priors. A publicly available multiple group version of the BILOG (Mislevy and Bock 1982) computer program called BIMAIN (Muraki and Bock 1987, 1991) has many of the same capabilities for dichotomously scored items only. Since the PARSCALE program was applied to dichotomously scored items in the ECLS-K vertical scaling, its estimation procedure is identical to the multiple group version of BILOG or BIMAIN. PARSCALE uses a marginal maximum likelihood estimation approach and thus does not estimate the individual ability scores when estimating the item parameters but assumes that the ability distribution is known for each subgroup. Thus, the posterior distribution of item parameters is proportional to the product of the likelihood of observing the item response vector, based on the data and conditional on the item parameters and subgroup membership, and the assumed prior ability distribution for that subgroup. More formally, the general model in terms of item-parameter estimation is the same as that used in NAEP and described in some detail by Yamamoto and Mazzeo (1992, p. 158) as follows:

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

In equation (3.2), $P(x_{j:g} \mid \theta, \beta)$ is the conditional probability of observing a response vector $x_{j:g}$ 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., $A_g(X_k) = f_g(X_k) / \Sigma_k f_g(X_k)$).

www.manaraa.com

Maximization of $L(\beta)$ is carried out by an application of an EM algorithm (Dempster, Laird, and Rubin 1977). When population densities are assumed known and held constant during estimation, the algorithm proceeds as follows. In the E step, provisional estimates of item parameters and the assumed multinomial probabilities are used to estimate expected sample sizes at each quadrature point for each group (denoted $\hat{N}_{gk}$), as well as over all groups (denoted $\hat{N}_k = \sum_g \hat{N}_{gk}$). These same provisional estimates are also used to estimate an expected frequency of correct responses at each quadrature point for each group (denoted $\hat{r}_{gik}$), and over all groups (denoted $\hat{r}_{ik} = \sum_g \hat{r}_{gik}$). In the M step, improved estimates of the item parameters, β, are obtained using maximum likelihood by treating the $\hat{N}_{gk}$ and $\hat{r}_{ik}$ as known, subject to any constraints associated with prior distributions specified for β.

The user of the multiple group version of PARSCALE has the option of fixing the priors on the ability distribution or allowing the posterior estimate to update the previous prior and combine with the data-based likelihood to arrive at a new set of posterior estimates after each major EM cycle. If one wishes to update on each cycle, one can continue to constrain the priors to be normal or their shape can be allowed to vary. The ECLS-K approach was to allow for updating the prior but with the normality assumption. The smoothing that came from the updated normal priors led to less jagged-looking ability distributions and did not tend to overfit the item parameters. Lack of fit in the item parameter distribution would simply be absorbed in the shape of the ability distribution if the updated ability distribution were allowed to take any shape. A similar procedure was used in estimating the item parameters in the National Adult Literacy Study (NALS) (Kirsch et al. 1993).

It should be remembered that the solution to equation 3.2 finds those item parameters that maximize the likelihood across all seven time points (the six 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 theta, were obtained from the C-Group conditioning program, which uses the Gaussian quadrature procedure. This procedure is virtually equivalent to conditioning (e.g., see Mislevy et al. 1992) on a set of "dummy" variables defining the ability subpopulation from which an observation comes. The one difference is that the group variances are not restricted to be equal as in the standard conditioning procedure.

Conditional independence is an assumption of all IRT models, but as Mislevy et al. (1992) point out, it is a strong assumption that is often violated in practice. However, if one thinks of IRT-based scores as a summarization of essentially the largest latent factor underlying a given item pool, then small

www.manaraa.com

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 graphs were inspected to ensure a good fit throughout the ability range. For each item, the empirical proportion correct in each round was computed, and compared with the model-based estimated proportion correct based on thetas for the same set of students, that is, the subset of students in the round who had received and responded to the item. Discrepancies between predicted and actual item proportion correct were reviewed for each round. No systematic over- or under-prediction was found for any round or for any type of item.

Tables B1 to B3 in appendix B list the IRT item parameters for the three subject areas. The items are sorted in ascending order of difficulty (the IRT "b" parameter). These tables also show the assessment versions in which the items appeared: one set of tests used for the first four rounds, fall- and spring-kindergarten and fall- and spring-first grade, with new versions used in third and fifth grades. Items that appeared in more than one assessment version served to link the scales across rounds (see section 5.1). Appendix B also shows the mean and standard deviation of the IRT ability estimate, theta, within each round. Bands marking two standard deviations below and above the theta mean illustrate the match of assessment difficulty to the range of student ability in each round. Tables C1 to C3 in appendix C show estimates of the proportion of correct responses to each item that would have been expected if all children had answered all of the items in the kindergarten through fifth-grade item pools at every round. Although each child answered only a small subset of the items each time, IRT ability estimates and item parameters make it possible to estimate performance on all of the items in the pool. In appendix D, tables D1 to D3 show the fit of the IRT model to the item response data. The IRT-estimated probability of a correct response was calculated for each item answered by each child. The average of these probabilities is equivalent to the estimated proportion correct predicted by the IRT model for each answered item. These estimates were compared with the actual proportion correct observed for the answered items. The tables in appendix D show the differences for each item (actual minus predicted), for all items used in each round. For nearly all items in nearly all rounds, these discrepancies were small, indicating good fit of the IRT model to the item response data.

3.3 Rating Scale Model

A generalization of the simple Rasch model (1960), the Rating Scale model (Wright and Masters 1982) was used to estimate the scores on the Academic Rating Scale (ARS) described in chapter

www.manaraa.com

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 power is constant across the items, and there is no guessing parameter. Applying Rasch models 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, Rasch models assume unidimensionality, that is, a single dimension is being measured.

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

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

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_j 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 is a subscript that varies between 0 and *m*.

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

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

www.manaraa.com

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.

3.3.1 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 (Normal Approximation Estimation Algorithm) is used. PROX assumes a normal distribution and does not take advantage of the ability of the simple Rasch model to calibrate measures independent of the sample characteristics (Wright and 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 item and person parameters, and indicate the replicability of the placement of the persons and items. The person reliability is analogous to Cronbach's alpha (table 7-1). Fit statistics are also provided for both persons and items (table 7-2). Both an information-weighted (infit) and an outlier sensitive (outfit) statistic are provided. The outfit mean square is sensitive to unexpected responses 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.

3.4 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

www.manaraa.com

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 (1989) point out, items so identified must still be interpreted in light of the intended meaning of the test scores before any conclusion of bias can be drawn. It is not entirely clear how the term item bias applies to academic achievement measures given to students with different patterns of exposure to content areas. For example, some students may be in schools where the third- through fifth-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, multiform 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. Although the M-H statistic looks at passing rates for two groups while controlling for total score, no assumption need be made about the shape of the total score distribution for either group. 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).

www.manaraa.com

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 fifth-grade assessment items for six sets of contrast groups: males (reference group) compared with females (focal group), and White children (reference group) compared with four other racial/ethnic groups: Black, Hispanic, Asian, and "Other." There were too few Native American and Multiracial children for DIF statistics to be evaluated separately for these groups. 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 fifth-grade assessment are discussed in chapter 4.

www.manaraa.com

This page is intentionally left blank.

www.manaraa.com

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

This chapter documents the direct cognitive test results for the fifth-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 five kindergarten through third-grade rounds are reviewed, to the extent that they are relevant to interpretation of fifth-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 may also differ from statistics reported in the users' manual. National estimates in this chapter are based on *all* children who had been tested at each round, using the corresponding cross-sectional weights, (C1CW0–C6CW0). Tables in the users' manual are based on the panel sample, that is, the subset of children who participated in all six rounds of data collection, and the longitudinal panel weight (C1_6SC0). The emphasis in this chapter is on the psychometric characteristics of the tests at each round, while the users' manual is designed to provide a reference for comparison with statistics obtained from secondary analyses, which may typically employ multiple rounds of data. Score statistics for all direct cognitive scores are presented in appendix A, with breakdowns by gender, race/ethnicity, socioeconomic status, and school type.

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

4.1 Types of Scores

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

www.manaraa.com

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), third-grade, or fifth-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 prereading skills, while the routing tests in third and fifth grades contained easy and difficult decoding words, understanding of words in context, 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- and fifth-grade routing tests were each based on different and more difficult sets of items. The third- and fifth-grade routing test number-right scores should *not* be compared with the kindergarten or first-grade routing test number-right scores, nor with each other.

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

Broad-based scores based on the full set of assessment items in reading, mathematics, and science were calculated using IRT procedures. The IRT scale scores estimate children's performance on

www.manaraa.com

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 186 scored questions in all of the first- and second-stage reading forms administered in all rounds, the 153 scored questions in all of the mathematics forms from all rounds, and the 92 third- and fifth-grade science items. (A small number of additional items was administered but not included in scale scores for reasons explained in sections 4.3 and 4.4.) These scores are not integers because they are probabilities of correct answers, summed over all items in the pools. (Scores for different subject areas are not comparable to each other because they are based on different numbers of questions, as well as content that is not necessarily equivalent in difficulty. That is, it would not be correct to assume that a child is doing better in reading than in mathematics because his or her IRT scale score is higher for reading than for mathematics.) A description of IRT methodology may be found in chapter 3. Chapter 5 contains a discussion of the application of IRT to creating longitudinal scores for ECLS-K.

Standardized scores (T-scores) provide norm-referenced measurements of achievement, that is, cross-sectional estimates of achievement *relative to the population as a whole*. A high mean T-score for a particular subgroup indicates that the group's performance is high in comparison with other groups. It does not represent mastery of a particular set of skills, only that the subgroup's mastery level is greater than a comparison group. Similarly, a change in mean T-scores over time reflects a change in the group's status with respect to other groups. In other words, T-scores provide information on *status compared with 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 fifth 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.

www.manaraa.com

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. 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 fifth 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 six rounds of data collection, less than 7 percent of reading response patterns and about 3 percent of mathematics assessment results failed to follow the expected hierarchical pattern. This does not necessarily indicate a different order of learning for these children; since most of the proficiency level items were multiple choice, many of these reversals may be due to children guessing.

www.manaraa.com

The nine reading and nine mathematics proficiency levels identified in the kindergarten through fifth-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 and fifth grades, children's data did not include all of the assessment items necessary to determine pass/fail for every proficiency level at each round of data collection. The missing information was not missing at random; it depended in part on children being routed to second-stage forms of varying difficulty within each round, and in part on the range of difficulty of the assessments at the different grade levels. In order to avoid bias due to the non-randomness of the missing proficiency level scores, imputation procedures were undertaken to fill in the missing information.

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

www.manaraa.com

correct answer were computed for each missing assessment item and used to assign an imputed right or wrong score to the item. These imputed responses were then aggregated in the same manner as actual responses to determine mastery at each of the missing levels. About 67 percent of the "highest level" scores in reading and about 80 percent in 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 alone, or by a combination of item responses and imputed scores. The highest proficiency level mastered implies that children demonstrated mastery of all lower levels and nonmastery of all higher levels. A zero score indicates nonmastery of the lowest proficiency level. Scores were excluded only if the actual or imputed mastery level data resulted in a reversal pattern as defined above. The highest proficiency level mastered scores do not necessarily correspond to an interval scale, so in analyzing the data, they should be treated as ordinal.

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

www.manaraa.com

are making their largest gains in achievement during a particular time interval. For example, subtracting the mathematics level 6 probability at third grade from the mathematics level 6 probability at fifth grade would indicate to what extent a student has advanced in mastery of place value 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 vast 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 in fifth grade, with well over 90 percent of children 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. In the early rounds, limited English proficiency was the primary reason for some children being excluded from the cognitive assessments; this was no longer a factor by fifth grade.

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 1/3 of 1 percent of testing sessions could not be completed, primarily because of scheduling difficulties or children's mental or physical limitations. Of the completed assessments, more than 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 Programs

www.manaraa.com

(IEPs). More details on accommodations provided during data collection can be found in *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Combined User's Manual for the ECLS-K Fifth-Grade Data Files and Electronic Codebooks* (NCES 2006–032) (Tourangeau et al. forthcoming); *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 Codebook* (NCES 2003–003) (Tourangeau et al. 2003); and *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) (Tourangeau et al. 2004). 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 6: School years 1998–99, 1999–2000, 2001–02, and 2003–04

Category	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Number of cases	19,045	19,884	5,253	16,684	14,383	11,298
Very low: Child doesn't try or attempt many items, even with encouragement	1.7	1.6	1.0	1.2	1.4	0.9
Low: Child frequently says "I don't know" without even trying, consistent encouragement needed	9.9	10.4	7.5	8.1	6.8	6.3
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 out the item.	48.5	44.5	44.9	39.5	33.7	34.2
High: Child tries or attempts every item, including some of the most difficult.	29.8	30.7	32.5	31.4	35.3	37.2
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	21.4
Very low + Low	11.6	11.9	8.5	9.3	8.3	7.2
Average + High + Very high	88.4	88.1	91.5	90.7	91.7	92.8

NOTE: Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. 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, spring 2002, and spring 2004.

www.manaraa.com

Table 4-2. Child's overall cooperation during the assessment, in percent: Rounds 1 through 6: School years 1998–99, 1999–2000, 2001–02, and 2003–04

Category	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Number of cases	19,046	19,884	5,253	16,684	14,383	11,298
Very uncooperative: Child repeatedly refuses to comply.	1.1	0.6	0.4	0.8	0.8	0.9
Uncooperative: Child complies at least 50 percent of the time.	2.7	2.0	1.5	1.3	0.9	0.4
Matter of fact: Child complies at least 75 percent of the time.	22.7	23.5	22.1	23.2	14.5	12.9
Cooperative: Child complies with most (80-90 percent) requests and directives.	53.2	49.6	49.9	43.5	44.1	43.2
Very cooperative: Child complies with all requests and directives in first request.	20.3	24.3	26.1	31.1	39.7	42.5
Very uncooperative + Uncooperative	3.8	2.6	1.9	2.2	1.7	1.3
Matter of fact + Cooperative + Very cooperative	96.2	97.4	98.1	97.8	98.3	98.7

NOTE: Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. 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, spring 2002, and spring 2004.

www.manaraa.com

Table 4-3. Child's overall attention level during the assessment, in percent: Rounds 1 through 6: School years 1998–99, 1999–2000, 2001–02, and 2003–04

Category	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Number of cases	19,046	19,884	5,253	16,684	14,383	11,298
Unable to attend: Child needs ongoing redirection to the task.	0.6	0.6	0.3	0.3	0.4	0.0
Difficulty attending: Child is distracted easily and often requires redirection.	13.6	11.4	8.0	9.4	9.0	6.5
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	29.4
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	36.1
Complete and full attention: Child is able to ignore any distractions.	11.5	16.3	18.7	22.5	25.5	28.0
Unable to attend + Difficulty attending	14.2	12.0	8.3	9.7	9.4	6.5
Attentive + Very attentive + Complete and full attention	85.8	88.0	91.7	90.3	90.6	93.5

NOTE: Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. 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, spring 2002, and spring 2004.

4.3 Reading Assessment

The fifth-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 26 items, 7 of which were based on a short reading selection. Three items tested understanding of vocabulary words in context. The remaining 16 items were decoding words, administered in ascending order of difficulty. Discontinue rules were in place for the routing test: when a child was not able to read a specified number of the decoding words in each progressively more difficult 4-item cluster, subsequent clusters were not administered. The score on the routing test was used to select one of three second-stage forms, of varying

www.manaraa.com

difficulty, each consisting of 4 (low and middle forms) or 5 (high form) reading passages, each with 4 to 8 associated questions. The low form also contained four individual word-in-context questions repeated from the earlier rounds.

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, and an additional set in round 6. 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. Scores were calculated only for children who attempted at least 10 items in the routing test and second-stage form combined. The line labeled "Too few items" refers to the number of children who did not attempt a sufficient number of reading items to generate a reliable score. This number is excluded from the "Total" line, which is the number of scoreable tests. Children who lacked sufficient English proficiency to pass the English language screening test, administered in rounds 1 through 4 only, were excluded from the reading assessment.

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- and fifth-grade assessments developed for rounds 5 and 6 were designed to route approximately 50 percent of children to the middle form, with the remaining children about evenly divided between the low and high forms.

www.manaraa.com

Table 4-4. Reading assessment: Samples and operating characteristics: Rounds 1 through 6: School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristics	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Total	17,630	18,944	5,054	16,340	14,286	11,267
Too few items	44	19	0	2	134	31
Number taking low form	13,355 (76%)	6,521 (34%)	1,062 (21%)	618 (4%)	3,540 (25%)	2,924 (26%)
Number taking middle form	3,620 (21%)	8,906 (47%)	2,334 (46%)	2,371 (15%)	8,032 (56%)	5,536 (49%)
Number taking high form	654 (4%)	3,517 (19%)	1,657 (33%)	13,351 (82%)	2,714 (19%)	2,807 (25%)
Percent perfect score routing test	.3	1.7	4.9	23.6	3.4	0.1
Percent perfect score low form	0.0	0.1	0.4	1.6	0.0	0.4
Percent perfect score middle form	0.0	0.0	0.0	0.0	0.0	0.0
Percent perfect score high form	0.0	0.2	0.0	0.0	0.0	0.5
Percent less than chance routing test	22.6	3.7	2.1	0.3	0.4	1.3
Percent less than chance low form	0.9	0.5	0.2	0.6	3.6	0.7
Percent less than chance middle form	0.5	0.3	0.1	0.1	0.2	0.2
Percent less than chance high form	0.5	1.7	2.3	0.4	0.0	0.0

NOTE: Rounds 1–4 used the same set of assessment forms; Round 5 and round 6 forms were different sets developed for third and fifth grades, respectively. Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. "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 may 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, spring 2002, and spring 2004.

www.manaraa.com

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 four correct answers on the routing and second-stage forms combined. The fifth-grade test forms were well matched to the ability levels of the tested children: only a fraction of 1 percent of test takers had a below-chance or perfect score on the routing and second-stage items 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) (Rock and Pollack 2002) 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 Codebook,* NCES 2002–149) (Tourangeau, Nord, et al. 2002). These items were not included in the third-grade reading forms because nearly all children had mastered them by the end of first grade.

www.manaraa.com

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

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

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

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

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

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

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

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

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

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

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

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

4.3.3 Reliabilities

Table 4-5 presents reliability statistics for the scores of the fifth-grade reading assessment. K-1 and third-grade reliabilities are included in the table for comparison purposes. In general, the more

www.manaraa.com

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 6: School years 1998–99, 1999–2000, 2001–02, and 2003–04

Reliability measure	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Alpha routing	.86	.88	.88	.86	.75	.88
Alpha low form	.69	.69	.71	.72	.83	.82
Alpha middle form	.70	.72	.74	.78	.84	.72
Alpha high form	.90	.88	.93	.92	.79	.76
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	.64
Split-half: Proficiency level 6	†	†	†	†	.48	.51
Split-half: Proficiency level 7	†	†	†	†	.48	.48
Split-half: Proficiency level 8	†	†	†	†	.63	.64
Split-half: Proficiency level 9	†	†	†	†	†	.40
Reliability of theta	.91	.93	.95	.96	.93	.94
Percent agreement of highest proficiency level mastered:						
Percent exact agreement	63	54	55	55	50	51
Percent exact + off by 1	96	94	94	95	95	95

† Not applicable.

NOTE: Statistics are unweighted. Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. Statistics are unweighted. Statistics for IRT-based scores (percent agreement and reliability of theta) may be different from those in earlier reports due to recalibration of longitudinal scales.

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

Internal consistency (alpha) coefficients for fifth grade are comparable to those obtained for K-1 and third grade. The pattern of alpha coefficients for the routing tests is at least in part due to the number of items. For tests with similar characteristics, a larger number of items will result in a higher alpha coefficient. The K-1 reading routing test had 20 items, with 15 items in third grade and 26 in fifth grade, and the resulting reliabilities follow the same pattern. The alpha coefficients for the second-stage forms in each round are generally lower than those for the routing test due to the restriction in range

www.manaraa.com

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

Split-half reliabilities were computed for the scores that are defined by clusters of items: the decoding score and the individual proficiency level scores. Each of these reliabilities is a transformation of the correlation of a subscore based on half of the items in the cluster with the score based on the other half. The decoding score was reported only for third and fifth grades, not for the earlier rounds. In the fifth-grade round, only three of the four items in this cluster were present in the assessment and the fourth item was imputed to produce a score, so a calculation of split-half reliability based on all items was not possible. Split-half reliabilities are presented for the individual proficiency level scores for 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 6–9 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. Error variance was estimated as the within-person variance of repeated estimates of theta, averaged over all data cases. The ratio of this number to the total variance (between-person variance of the posterior mean) is the estimated proportion of total variance that is error variance, and 1 minus the proportion is the estimate of true variance that is reported as the reliability of theta. This reliability index differs from the information function primarily in that it is a single estimate for the whole set of scores, rather than a function evaluated at each point along the continuum. This is the most appropriate estimate of the reliability of the assessment since it reflects the internal consistency of performance on the combined first- and second-stage sections, and for the full range of variance found in

www.manaraa.com

the sample as a whole. The reliability of theta applies to the scale scores and proficiency probabilities as well, since these scores are nonlinear transformations of the thetas that do not affect rank orderings.

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

The definition of reliability–consistency of measurement under different circumstances–suggested an appropriate way to assess the reliability of the "highest proficiency level mastered" score. The score denoting the highest level mastered reduces the series of pass/fail scores on the hierarchical set of proficiency levels to a single score. For example, a 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 186 test items. Four additional items, consisting of difficult decoding words, were used for the purpose of calibrating IRT ability, but deleted from the score scale to bring the representation of content strands more closely into alignment with the framework specifications. One tested item was deleted from scoring due to differential item functioning (DIF) (see section 4.3.5). The IRT procedures described earlier allowed the scale score estimates to be computed based on the subset of questions actually administered to each child at each round. As the assessments progressed from kindergarten through fifth

www.manaraa.com

grade, more and more of the test items relied on comprehension of reading passages. Inspection of the reading scale score means by round shows an accelerated rate of growth between fall and spring of first grade, round 3 to round 4, and much larger gains between first and third grade, round 4 to round 5. These gains correspond to the times when children would be mastering basic technical reading skills, and then later, acquiring the ability to derive meaning from what they read. The greater variability in reading performance in the later rounds, compared with kindergarten and fall first grade, can be interpreted as an increase in the reading skills gap between low and high achievers. 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 6: School years 1998–99, 1999–2000, 2001–02, and 2003–04

	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Scale score mean	29.0	40.1	46.8	70.2	116.1	136.7
Scale score standard deviation	9.8	13.4	17.2	22.4	25.6	24.3

NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3A5W0, C4A3W0, C5CW0, C6CW0). Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. Estimates for kindergarten through fifth grade have been put on a common scale to support comparisons. The range of values: 0–186.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

4.3.5 Differential Item Functioning

Section 3.4 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 fifth-grade reading items. The largest number of C-DIF[1] items was found for performance comparisons of White versus 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 fifth-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 fifth-grade assessments. Second, many of the Asian children came from a language minority background. Two of the three items on which Asian children performed relatively better than expected were difficult decoding items, while the four questions that were relatively harder for Asian children involved inferences based on stories. (Compare these numbers

[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 experiences that effects sizes of 1.5 and higher have practical significance. Effect sizes of this magnitude that are statistically significant are labeled with a "C."

www.manaraa.com

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 either group alone, and, for the two groups combined as "other," no C-DIF items were found.

Table 4-7. Reading assessment: Differential item functioning, fifth grade: School year 2003–04

Reference group: Focal group:	Male Female	White Black	White Hispanic	White Asian	White Other
Number of C-DIF[1] items favoring focal group	0	0	1	3	0
Number of C-DIF items favoring reference group	3	1	1	4	0

[1] DIF having an effect size of 1.5 or greater, hence statistically significant.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

It should be kept in mind that there were 94 reading items in the fifth-grade reading assessment forms and five 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. However, one item was judged to be differentially more difficult for Asian children because of cultural considerations and was not scored.

4.4 Mathematics Assessment

The fifth grade mathematics framework specifications were identical to those for third grade, in terms of percentages of items in each content strand for the whole item pool, and quite similar to those for the kindergarten and first-grade rounds. 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 and geometry items. Greater emphasis was placed on problem solving in fifth grade compared with the earlier rounds. Children began the mathematics assessment with a routing test of 18 items. The score on the routing test was used to select one of three second-stage forms, of varying difficulty, each consisting of 18 (low and middle forms) or 19 (high form) items.

www.manaraa.com

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 in those rounds. More advanced sets of assessment forms, entirely in English, were developed for third and fifth grades. Scores were calculated only for children who attempted at least 10 items in the routing test and second-stage form combined.

The fifth-grade assessment developed for round 6 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 fifth-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.

www.manaraa.com

Table 4-8. Mathematics assessment: samples and operating characteristics, rounds 1 through 6: School years 1998–99, 1999-2000, and 2001–02

Characteristics	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Total	18,641	19,657	5,226	16,647	14,380	11,276
Too few items	21	15	0	2	29	22
Number taking low form	14,380 (77%)	8,444 (43%)	1,353 (26%)	1,097 (7%)	4,229 (29%)	4,023 (36%)
Number taking middle form	3,123 (17%)	6,169 (31%)	1,521 (29%)	2,317 (14%)	5,344 (37%)	3,842 (34%)
Number taking high form	1,136 (6%)	5,042 (26%)	2,351 (45%)	13,233 (79%)	4,804 (33%)	3,410 (30%)
Percent perfect score routing test	0.1	0.4	1.5	7.9	1.6	1.8
Percent perfect score low form	0.1	0.4	1.0	2.5	0.0	0.6
Percent perfect score middle form	0.0	0.0	0.0	0.3	0.1	0.0
Percent perfect score high form	0.0	0.0	0.0	0.1	0.0	0.2
Percent less than chance routing test	15.3	3.1	1.6	0.3	1.3	1.5
Percent less than chance low form	0.9	0.3	0.1	0.3	0.3	0.4
Percent less than chance middle form	0.1	0.0	0.0	0.0	0.1	0.1
Percent less than chance high form	0.1	0.0	0.0	0.0	0.1	0.7

NOTE: Rounds 1–4 used the same set of assessment forms; rounds 5 and 6 forms were different sets developed for third and fifth grades. Some children in rounds 1–4 received a Spanish translation of the mathematics assessment; in rounds 5 and 6, all assessments were in English. Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. "Too few items" refers to the number of children who did not attempt a sufficient number of mathematics items to generate a reliable score. Percentages are unweighted. 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, spring 2002, and spring 2004.

www.manaraa.com

4.4.2 Scores Unique to the Mathematics Assessment: Proficiency Levels

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

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

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

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

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

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

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

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

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

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

As was the case for reading, the test items on which the mathematics proficiency levels were defined were not used in all rounds of data collection, but only in grades for which their difficulty was appropriate. Levels 1–3 items appeared only in the K-1 assessments, level 4 in K-1 and third grades, level 5 in all rounds, levels 6–7 in third and fifth grades, and levels 8 and 9 in fifth grade only. (One item in each of the two highest proficiency levels had been present in the third grade test, but without the remaining three, third grade pass/fail scores for the levels could not be computed.) IRT procedures described in sections 3.2 and 5.2 were used to obtain probability estimates for all levels at all rounds so that longitudinal gains in specific skills could be measured.

www.manaraa.com

4.4.3 Reliabilities

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

Table 4-9. Mathematics assessment reliabilities, rounds 1 through 6: School years 1998–99, 1999–2000, 2001–02, and 2003–04

Reliability measure	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Alpha routing	.78	.81	.83	.80	.86	.88
Alpha low form	.70	.66	.66	.71	.77	.78
Alpha middle form	.66	.67	.66	.66	.72	.58
Alpha high form	.80	.80	.83	.82	.73	.75
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	.64
Split-half: Proficiency level 6	†	†	†	†	†	.78
Split-half: Proficiency level 7	†	†	†	†	.43	.68
Split-half: Proficiency level 8	†	†	†	†	†	.56
Split-half: Proficiency level 9	†	†	†	†	†	.48
Reliability of theta	.89	.91	.92	.92	.94	.94
Percent agreement of highest proficiency level mastered:						
Percent exact agreement	54	51	52	57	56	55
Percent exact + off by 1	97	95	96	97	97	97

† Not applicable.
NOTE: Statistics are unweighted. Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. The four test items for mathematics proficiency level 6 did not all appear in the same test form in third grade, so no complete data cases were available for evaluation of split half reliability. Statistics for IRT-based scores (percent agreement and reliability of theta) may be different from those in earlier reports due to recalibration of longitudinal scales.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

All other things being equal (e.g., the psychometric quality of test items), internal consistency coefficients tend to be higher when tests are longer and lower when the ability range of the test takers is restricted. The internal consistency (alpha) coefficients for the third- and fifth-grade mathematics routing tests were slightly higher than that of the K-1 forms, probably partly due to a slightly longer test (17 and 18 items in third and fifth grades, respectively, vs. 16 items in K-1), and partly because of greater variability in the mathematics achievement of third- and fifth-graders compared with

www.manaraa.com

earlier rounds. The fifth-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- and fifth-grade tests all had about the same number of items in each second stage form, and similar alphas. The reliabilities of the second-stage forms are presented for the sake of completeness, although scores on the second-stage forms are not reported separately.

Split-half reliabilities are shown in the table for the items present at each round: levels 1–3 items were present in the K-1 mathematics assessment only, level 4 in K-1 and third-grade, level 5 in all rounds, levels 6 and 7 in third and fifth grades, and levels 8 and 9 only in the fifth-grade forms. There is no split-half reliability presented for proficiency level 6 in third grade because the items on which it is based did not all appear in the same test form, so no complete data cases were available for evaluation of the reliability. The kindergarten and first-grade split-half reliabilities for levels 1 through 5 were substantially lower than for the corresponding levels in the reading test. While the sets of reading items in each of the lowest proficiency levels were essentially replicates of the same task, the mathematics sets were not as homogeneous with respect to content and skill demands. The greater heterogeneity for the mathematics sets may have contributed to their lower split-half reliabilities. Both alpha coefficients and split-half reliabilities tend to be underestimates of "true" reliability, and this tendency may be accentuated by greater diversity of content. The relatively low split-half reliabilities for mathematics proficiency levels 8 and 9 in fifth grade are a consequence of their placement only in the high level form, resulting in restriction in the range of ability of children taking these items.

Similar to the reading test, the reliabilities of the third- and fifth-grade theta scores were in the mid .90s. Reliabilities for the K-1 rounds were lower than had been reported for earlier versions, because the score scale extended through fifth grade increasingly emphasized problem solving. The reliability of theta applies to the scale scores and proficiency probabilities as well, since these scores are nonlinear transformations of the thetas that do not affect rank orderings.

The percentages of agreement between methods in determining the highest mathematics proficiency level mastered were comparable to those for reading, both for percentage of exact agreement, and percentage of agreement within one level. The greater homogeneity of the reading items for the low compared with high proficiency levels resulted in percent agreement of highest level that tended to go down in the later rounds. Conversely, percent agreement for mathematics, with greater heterogeneity in

www.manaraa.com

the *early* rounds, tended to go up. See section 4.3.3 for a detailed explanation of how this score was computed and evaluated.

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 153 mathematics items in the pool; that is, all items that appeared in any of the K-1, third-grade, and/or fifth-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, leveling off in fifth 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 6: School years 1998–99, 1999–2000, 2001–02, and 2003–04

Item	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Scale score mean	22.4	32.4	39.4	56.6	90.5	111.2
Scale score standard deviation	8.7	11.4	13.7	17.0	21.9	22.4

NOTE: Table estimates are based on cross sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. Estimates for kindergarten through fifth grade have been put on a common scale to support comparisons. The range of values: 0–153.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

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

4.4.5 Differential Item Functioning

Table 4-11 presents counts of the C-DIF items for the fifth-grade mathematics forms. There were insufficient numbers of Native American and multiracial children in the sample for DIF statistics to be computed for either group alone, and for the two groups combined as "other" no C-DIF items were found. All C-DIF mathematics items were reviewed and found to be relevant to the construct being

www.manaraa.com

measured by the assessment, and all were retained for scoring. See section 3.4 for an explanation of DIF procedures.

Table 4-11. Mathematics assessment: Differential item functioning, fifth grade: School year 2003–04

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

[1] DIF having an effect size of 1.5 or greater, hence statistically significant.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

4.5 Science Assessment

The fifth-grade science assessment consisted of a 21-item routing test followed by low, middle, and high difficulty second stage forms of 15, 17, and 14 items, respectively. Content of the science questions was approximately equally divided among life science, earth science, and physical science strands. The science assessment was first added to the ECLS-K cognitive battery in third grade; thus the longitudinal score scale spans only third to fifth grades.

4.5.1 Samples and Operating Characteristics

Table 4-12 presents sample counts and operating characteristics of the fifth-grade science forms. Scores were calculated only for children who attempted at least 10 items.

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

www.manaraa.com

Table 4-12. Science assessment: Samples and operating characteristics, rounds 5 and 6: School years 2001–02 and 2003–04

Characteristics	Round 5	Round 6
Total	14,357	11,273
Too few items	41	25
Number taking low form	4,199 (29%)	1,432 (13%)
Number taking middle form	7,204 (50%)	4,626 (41%)
Number taking high form	2,952 (21%)	5,210 (46%)
Percent perfect score routing test	1.5	1.1
Percent perfect score low form	0.3	0.0
Percent perfect score middle form	0.0	0.1
Percent perfect score high form	0.0	0.1
Percent less than chance routing test	4.7	1.1
Percent less than chance low form	1.7	4.9
Percent less than chance middle form	0.6	4.0
Percent less than chance high form	0.8	9.9

NOTE: No science assessment was conducted in rounds 1–4. The round 5 and round 6 assessments were developed for third and fifth grades. Percentages are unweighted. Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. "Too few items" refers to the number of children who did not attempt a sufficient number of science items to generate a reliable score. 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), spring 2002 and spring 2004.

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 21 routing form items of the fifth-grade science assessment tapped a range of basic concepts, with 7 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;

www.manaraa.com

- **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.

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

4.5.3 Reliabilities

Table 4-13 presents reliability coefficients for the third- and fifth-grade science assessments. Alpha coefficients for the routing test and second-stage forms are somewhat lower than those for reading and mathematics because the science assessment had fewer items in the second-stage forms. This is especially true for the fifth-grade science assessment, in which the routing test was lengthened to 21 items (from 15 in third grade) and the second-stage forms shortened to 14 to 17 items (from 20 in third grade) in order that the items designated for the three science cluster scores would be administered to all children. As a result, the alpha coefficient is higher for the routing test, and lower for the second-stage forms, than was the case in third grade. 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.

www.manaraa.com

Table 4-13. Science assessment reliabilities, rounds 5 and 6: School years 2001–02 and 2003–04

Reliability measure	Round 5	Round 6
Alpha routing	.75	.79
Alpha low form	.70	.54
Alpha middle form	.61	.63
Alpha high form	.60	.48
Split-half: Life Science 5 item cluster	.59	.59
Split-half: Physical Science 5 item cluster	.49	.41
Split-half: Earth Science 5 item cluster	.46	.52
Split-half: Life Science 7 item cluster	†	.64
Split-half: Physical Science 7 item cluster	†	.43
Split-half: Earth Science 7 item cluster	†	.62
Reliability of theta	.88	.87

NOTE: Statistics are unweighted. Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002 and spring 2004.

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

4.5.4 Score Statistics

Third- and fifth-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 92 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.

www.manaraa.com

Table 4-14. Science scale score mean and standard deviation, rounds 5 and 6: School years 2001–02 and 2003–04

Item	Round 5	Round 6
Scale score mean	43.5	56.1
Scale score standard deviation	14.1	14.9

NOTE: Table estimates are based on cross-sectional weights within each round (C5CW0 and C6CW0). Approximately 90 percent of the round 6 children were in fifth grade during the 2003–04 school year, 9 percent were in fourth grade, and about 1 percent were in third or other grades. Estimates for third through fifth grade have been put on a common scale to support comparisons.. The range of values is 0–92.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002 and spring 2004.

4.5.5 Differential Item Functioning

Table 4-15 summarizes the results of the DIF analysis of the fifth-grade science items. Only two items were identified as having C-DIF, and one of them favored the focal group (Hispanics). There were too few Native American and multiracial children in the sample for DIF to be evaluated for these children. No C-DIF was found for these two groups combined. The 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, fifth grade: School year 2003–04

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

[1] DIF having an effect size of 1.5 or greater, hence statistically significant.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002 and spring 2004.

Section 3.4 explains the DIF procedures used for identifying test items that perform differentially for population subgroups.

www.manaraa.com

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-fifth grade, and of science skills from spring-third grade to spring-fifth grade. The development of the longitudinal scales, including analysis of common items, will be described. Evidence supporting the validity of the measures will be presented. The final section of the chapter will focus on applications: choosing the appropriate scores for analysis and interpreting gain statistics.

5.1 Development of the K-1-3-5 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- and fifth-graders. Data from a small sample of second-graders was included to support the development of the scales by bridging the anticipated gap in ability between first and third grades. The link between the assessment forms used in different rounds relied on the presence of common items shared by successive test forms.

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

www.manaraa.com

5.1.1 Second-Grade Bridge Study

Chapter 2, section 2.1.5 of the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K)*, *Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05) (Rock and Pollock 2002) documents the gap in ability levels that was anticipated due to the absence of the second-grade data collection from the longitudinal design. Without any second-grade data, the accuracy of measurement of cognitive gains from first to third grade might have been compromised. Many of the cognitive test items linking the kindergarten through first-grade (K-1) assessments with the third-grade forms were too hard for most first-graders, and too easy for most third-graders. Stable estimates of item parameters necessary for establishing the longitudinal scale require that there be substantial numbers of test takers whose ability levels match the difficulty of the linking items. These test takers did not need to be part of the ECLS-K longitudinal cohort. They needed only to have ability levels in the range where the ECLS-K longitudinal sample data might be sparse, and to take sets of cognitive test items that included the items designed to link the first- and third-grade rounds. Section 5.1 of the above-referenced report describes in detail the collection of reading and mathematics data for a sample of approximately 900 second-graders who were not part of the ECLS-K longitudinal sample. It documents the characteristics of the second-grade bridge sample, and shows how the data were used to supplement the longitudinal sample data in establishing vertical scales for measurement of gain. Since the purpose of the bridge sample was to obtain data on the performance of the assessment items, rather than track the progress of the children themselves, their assessment scores are not included in released data files.

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

5.1.2 Evaluating Common Items

Linking score scales across grades required not only overlapping ability distributions, but also overlapping test forms. The longitudinal score scales relied on common items that were present in more than one set of assessment forms. These common items permitted the development of a vertical scale suitable for measuring gains in the elementary years. Table 5-1 shows the number of items in each subject area shared by more than one set of assessment forms, as well as the number that appeared in only

www.manaraa.com

one set. Within rounds, the score scale was supported by items taken by all students within the round (the 12 to 25 items on the routing tests) as well as smaller numbers of items overlapping two or all three second-stage forms.

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

Assessment versions	Reading	Mathematics	Science
Total item pool	186	153	92
Common items (total)	69	40	27
K-1 and third grade	13	9	†
First-grade supplement and fifth grade	2	0	†
Third and fifth grade	45	27	27
K-1 (or first grade supplement), third and fifth grade	9	4	†
Unique items (total)	117	113	65
K-1 version only	60	50	†
First-grade supplement only	8	†	†
Third grade only	16	34	35
Fifth grade only	33	29	30

† Not applicable.
NOTE: Four additional reading items were deleted from the scale scores to bring the content representation into closer alignment with framework specifications.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

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 three versions of the assessments (K-1, third grade, and fifth grade), it was still possible for the items to function differently. Of course, it would be expected that performance on the items would improve as children advance through school and gain skills, and gains in the probability of a correct answer would be observed. However, the *relative* difficulty of items in the context of the whole assessment should be maintained for the common items used to anchor the scale. For example, an item "X" based on content that had not yet been introduced could, in 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 and fifth grades, 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.

www.manaraa.com

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

Tables 5-2 through 5-4 compare the actual with the predicted proportion correct for each of the reading, mathematics, and science items used in more than one assessment version, based on the children who answered each of the items in each round of data collection. Note that the comparisons of observed vs. predicted percent correct for each question can be carried out *only for children who answered the question*. Many questions appeared in only one or two second-stage forms within a grade, or after a discontinue point in the routing test. Thus most of the items were answered by only a subset of children tested in each round. The statistics shown in tables 5-2 through 5-4 do not represent the difficulty of the items, but rather the fit of the IRT model to the data, evaluated on the basis of comparisons of actual and predicted responses for all items answered.

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

www.manaraa.com

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

Item	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
RUNS	K1,3	0.00	0.00	-0.01	0.03	-0.01	†
WENT	K1,3	0.00	0.00	-0.01	0.02	-0.01	†
DOWN	K1,3	0.00	-0.02	-0.01	0.03	-0.02	†
JEEP	K1,3	0.01	0.00	-0.01	0.02	-0.07	†
QUIET	K1,3	†	†	0.01	0.01	-0.01	†
RAGE	K1,3	0.06	0.02	-0.01	-0.01	0.03	†
TOIL	K1,3	0.06	0.02	0.01	-0.01	0.04	†
CORNER	K1,3	0.03	0.00	-0.01	-0.01	0.03	†
REQUIRE	K1,3	†	†	0.02	0.00	0.00	†
CAPTURE	K1,3	0.05	0.01	-0.02	0.00	0.01	†
WEB	K1,3	0.04	0.00	-0.01	0.00	0.01	†
STRANDS	K1,3	0.03	0.01	-0.01	0.00	-0.01	†
AMBITIO	K1,3	†	†	0.03	0.07	0.00	†
WAGES	K1,5	†	†	0.03	-0.03	†	0.00
ALIGNMNT	K1,5	†	†	-0.03	-0.11	†	0.01
RDLETR	3,5	†	†	†	†	0.00	0.01
RDMARIAB	3,5	†	†	†	†	0.00	0.00
RDGROSR	3,5	†	†	†	†	0.01	-0.01
RDLIKE	3,5	†	†	†	†	0.00	0.04
RDTIME	3,5	†	†	†	†	0.00	0.02
RDENDR	3,5	†	†	†	†	-0.01	0.00
RDFEELSR	3,5	†	†	†	†	0.00	0.00
RDSAMER	3,5	†	†	†	†	-0.01	0.00
RDGEORGR	3,5	†	†	†	†	0.02	-0.01
RDTANZAR	3,5	†	†	†	†	0.02	-0.02
RDDOCR	3,5	†	†	†	†	0.00	-0.01
RDSISR	3,5	†	†	†	†	0.00	-0.01
RDSTORY	3,5	†	†	†	†	0.01	-0.04
RDWAY	3,5	†	†	†	†	-0.02	0.08
RDKNIGHT	3,5	†	†	†	†	0.00	-0.03
RDJAMEDR	3,5	†	†	†	†	0.01	-0.02
RDCLUER	3,5	†	†	†	†	0.01	-0.01
RDBOWY	3,5	†	†	†	†	0.01	-0.01
RDTRAINY	3,5	†	†	†	†	0.01	-0.01
RDSUPRIR	3,5	†	†	†	†	-0.01	0.02
RDTEARB	3,5	†	†	†	†	-0.11	0.02
RDSAFER	3,5	†	†	†	†	0.00	0.02
RDBAKEDB	3,5	†	†	†	†	-0.02	0.01
RDTHREEB	3,5	†	†	†	†	-0.02	0.00

See notes at end of table.

www.manaraa.com

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

Item	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
RDMOVEBY	3,5	†	†	†	†	0.01	-0.01
RDLIKER	3,5	†	†	†	†	-0.01	0.04
RDDOMEST	3,5	†	†	†	†	-0.01	0.01
RDDIFFR	3,5	†	†	†	†	0.04	-0.04
RDINFLUB	3,5	†	†	†	†	-0.03	0.01
RDPROBLY	3,5	†	†	†	†	0.04	-0.04
RDBRETY	3,5	†	†	†	†	0.04	-0.04
RDJOSHB	3,5	†	†	†	†	0.08	-0.03
RDRACHLB	3,5	†	†	†	†	0.06	-0.02
RDTHEMEB	3,5	†	†	†	†	0.04	-0.02
RDMICROB	3,5	†	†	†	†	-0.01	0.01
RDSOLVEY	3,5	†	†	†	†	0.02	-0.02
RDPERSNB	3,5	†	†	†	†	0.04	-0.01
RDHELPY	3,5	†	†	†	†	0.02	-0.02
RDCOMPRB	3,5	†	†	†	†	0.01	0.00
RDGUESS	3,5	†	†	†	†	0.01	0.00
RDHOAXB	3,5	†	†	†	†	-0.02	0.03
RDCROPB	3,5	†	†	†	†	-0.02	0.03
DCIRCLB	3,5	†	†	†	†	0.02	-0.01
RDVORTXB	3,5	†	†	†	†	-0.01	0.02
RDWAGON	3,5	†	†	†	†	0.04	-0.01
BACKPACK	K1,3,5	0.03	0.02	0.02	-0.01	0.00	0.02
LISTEN	K1,3,5	0.04	0.01	0.01	0.00	0.00	-0.02
RIDEBIKE	K1,3,5	0.07	0.03	0.00	-0.01	0.00	0.00
SIZES	K1,3,5	0.03	0.03	0.01	-0.01	0.00	-0.02
THROUGH	K1,3,5	0.06	0.00	0.01	0.02	0.01	-0.03
WTLESS	K1,3,5	†	†	0.00	-0.02	0.01	0.01
MOISTURE	K1,3,5	†	†	0.00	-0.03	-0.01	0.01
CRITCISM	K1,3,5	†	†	0.01	-0.08	-0.03	0.05
PREFRNCE	K1,3,5	†	†	0.12	0.09	0.00	-0.02

† Not applicable.
NOTE: Positive numbers correspond to actual proportion correct that is higher than predicted by the IRT model, and negative numbers to actual proportion correct that is lower than predicted.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

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

Item	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
2+5MARBL	K1,3	0.02	0.01	0.04	0.01	-0.16	†
12 BY 2S	K1,3	0.01	0.00	-0.03	0.01	-0.01	†
3+7PENNY	K1,3	0.02	-0.01	-0.01	0.01	-0.01	†
51015_25	K1,3	-0.02	-0.01	-0.04	0.03	0.06	†
4+4-2	K1,3	-0.01	0.01	0.02	0.01	-0.05	†
HOWMANY$	K1,3	0.05	0.02	0.02	-0.04	0.07	†
12-? PEN	K1,3	0.06	0.04	0.03	-0.03	0.02	†
HEADSUP	K1,3	0.05	0.00	0.01	-0.03	0.06	†
GOALS	K1,3	0.00	0.00	0.00	0.00	0.00	†
CUBES10	3,5	†	†	†	†	0.00	0.00
NEXT78	3,5	†	†	†	†	-0.01	0.01
DO_ADD4	3,5	†	†	†	†	0.02	-0.04
TIME1030	3,5	†	†	†	†	0.00	0.01
NUMBER60	3,5	†	†	†	†	0.00	-0.01
CUBESIDE	3,5	†	†	†	†	-0.01	0.01
NEXT120	3,5	†	†	†	†	-0.01	0.00
CHART_64	3,5	†	†	†	†	0.00	-0.01
BOX_700	3,5	†	†	†	†	0.00	-0.01
SPOONS	3,5	†	†	†	†	0.00	-0.01
COLORSYM	3,5	†	†	†	†	-0.06	0.03
PAGES78	3,5	†	†	†	†	0.01	-0.02
A568214K	3,5	†	†	†	†	-0.01	0.02
CHARGE_5	3,5	†	†	†	†	-0.03	0.03
MARIA310	3,5	†	†	†	†	0.03	-0.02
CARDS579	3,5	†	†	†	†	-0.03	0.02
PAIR_100	3,5	†	†	†	†	-0.02	0.04
GREW4_	3,5	†	†	†	†	0.02	-0.03
LOUISA13	3,5	†	†	†	†	0.03	-0.02
MIN_BLOW	3,5	†	†	†	†	0.04	-0.03
TALL75_	3,5	†	†	†	†	-0.05	0.02
MARBLES	3,5	†	†	†	†	0.00	0.03
BANKER_	3,5	†	†	†	†	0.00	-0.01
MARK_DOT	3,5	†	†	†	†	0.01	-0.01
EDGECUBE	3,5	†	†	†	†	0.00	0.00
SAMEFRAC	3,5	†	†	†	†	-0.01	0.01
TILESCOV	3,5	†	†	†	†	0.01	0.00

See notes at end of table.

www.manaraa.com

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

Item	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
A13_79	K1,3,5	-0.02	-0.03	-0.03	0.04	0.00	-0.02
COST_10	K1,3,5	0.05	0.02	0.04	-0.03	0.00	-0.02
CARS15_5	K1,3,5	0.04	0.02	0.02	-0.01	-0.01	-0.02
CANDY8_2	K1,3,5	0.02	0.02	0.04	-0.03	0.01	0.01

† Not applicable.
NOTE: Positive numbers correspond to actual proportion correct that is higher than predicted by the IRT model, and negative numbers to actual proportion correct that is lower than predicted.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

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

Item	Used in grades	Round 5	Round 6
ROUIMM	3,5	0.00	0.00
RWINGS	3,5	0.02	-0.07
ROUFRZ	3,5	-0.01	0.00
ROUTAP	3,5	-0.02	0.02
ROUJUN	3,5	0.00	-0.01
ROUERT	3,5	0.01	-0.02
ROUBRN	3,5	0.00	-0.01
RHEART	3,5	0.00	0.01
ROUJAR	3,5	0.00	-0.01
ROUSRF	3,5	0.00	0.00
RDESRT	3,5	0.01	-0.02
YTHEMT	3,5	0.02	-0.02
YMOON	3,5	0.03	-0.03
ROUSOL	3,5	-0.04	0.05
YBEES	3,5	0.01	-0.03
ROUBLB	3,5	-0.01	0.02
ROUMTN	3,5	-0.02	0.03
ROUGRT	3,5	0.00	0.01
ROUMCE	3,5	0.01	-0.01
ROUFLY	3,5	0.00	0.01
BSOUND	3,5	-0.05	0.03
ROUSHD	3,5	0.04	-0.05
BPLNT2	3,5	-0.02	0.01
BPLANT	3,5	0.08	-0.02
BSLIDE	3,5	-0.06	0.02
BSOIL	3,5	-0.03	0.02
BMAMML	3,5	0.05	-0.02

NOTE: Positive numbers correspond to actual proportion correct that is higher than predicted by the IRT model, and negative numbers to actual proportion correct that is lower than predicted.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002 and spring 2004.

www.manaraa.com

5.1.3 IRT Calibration and Scoring

IRT calibration was carried out using the PARSCALE program as described in chapter 3. One reading item was deleted from the item pool because of differential item functioning (DIF) for a population subgroup (see section 4.3.5). The estimation of reading item parameters and student abilities was based on the remaining 190 unique items that appeared in all forms of the reading assessments, with the 69 items common to two or more of the assessment versions serving to anchor the scale. Four of the 190 items were deleted from the final scale scores so that the scale would be more closely aligned with framework specifications, leaving 186 items in the final reading scale. No mathematics items were deleted because of differential functioning in grade 5 (two had been deleted for this reason in earlier rounds). The K-1, third-grade, and fifth-grade mathematics scale is based on 153 unique mathematics items in all assessment forms, including 40 common to more than one version of the assessment. The science scale is based on 92 unique items in third and fifth grades, including 27 common to both rounds. For each item, the IRT calibration resulted in a set of three item parameters that define a logistic function associated with the item. The height of the function at any point along an ability range corresponds to the estimated probability of a correct answer on the item for a person at that ability level. The tables in appendix B show the item parameters, in ascending order of difficulty (IRT "b" parameter).

Each of the rounds of data collection, kindergarten through fifth grade (plus the bridge sample), was treated as a separate subpopulation with its own ability distribution for the purpose of IRT calibration. This feature of PARSCALE and other Bayesian approaches to IRT provides for an empirically based shrinkage toward subpopulation means for extreme ability estimates, low and high. This shrinkage is particularly important for a longitudinal study, where the focus is on measuring gain and it is important to avoid floor and ceiling effects. See section 3.2.1 for additional details. Table 5-5 presents theta (ability) means and standard deviations for the subpopulations of the reading, mathematics, and science calibrations. The theta estimates are standardized to mean = 0.0 and standard deviation = 1.0 for all rounds combined.

www.manaraa.com

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

Round	Reading		Mathematics		Science	
	Mean	SD[1]	Mean	SD	Mean	SD
All rounds combined	0.00	1.00	0.00	1.00	†	†
Round 1 (fall-kindergarten)	-1.18	0.51	-1.12	0.51	†	†
Round 2 (spring-kindergarten)	-0.60	0.50	-0.59	0.49	†	†
Round 3 (fall-first grade)	-0.36	0.51	-0.32	0.49	†	†
Round 4 (spring-first grade)	0.25	0.46	0.22	0.44	†	†
Second grade bridge sample	0.83	0.29	0.69	0.31	†	†
Round 5 (spring-third grade)	0.99	0.35	0.94	0.41	-0.38	0.86
Round 6 (spring-fifth grade)	1.30	0.35	1.39	0.46	0.42	0.87

† Not applicable.
[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, six data collection rounds, 1998–99, 1999–2000, 2001-02, and 2003-04, plus bridge sample 2001–02.

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 186 reading items, all 153 mathematics items, or all 92 science items. The T-scores in the database are theta estimates transformed to a metric of mean = 50.0, standard deviation = 10.0 within each round, using cross sectional sample weights.

Proficiency scores required an additional IRT calibration step. Section 4.1.4 describes the selection of a hierarchical series of mastery levels in reading, and another series in mathematics, marked by clusters of four items at each level. Nine such levels were defined in each subject, based on items from the K-1, third-grade, and fifth-grade assessments. Children were judged to have passed a level (score = 1) if they answered at least three of the four items correctly, and to have failed if at least two wrong answers were given (score = 0). Children with fewer than three right or two wrong answers (because they omitted items, or because the items defining a particular level were not included in the assessment forms they received) were not scored for the purpose of IRT calibration. The proportion of omitted responses in all subjects in all rounds was negligible, so nearly all children had pass or fail scores on the proficiency

www.manaraa.com

levels whose items were administered to them. After the initial PARSCALE estimates of item parameters and abilities were obtained, parameters for the proficiency levels were estimated. Ability levels were held constant, and the proficiency level clusters (scored as right, wrong, or not administered) were treated as items for estimating item parameters. In essence, this resulted in prediction of mastery level proficiency from estimates of ability levels derived from all items administered to each child. Extremely close fits of the logistic functions to the proportion correct from item-response-based cluster scores (1 or 0) were observed for all levels in all rounds, for both reading and mathematics.

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

The parameters for the reading and mathematics proficiency levels are shown in table 5-6. The very high "a" parameters are consistent with the assumption that 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 5-6. IRT parameters for reading and mathematics proficiency levels, based on items from kindergarten, first-grade, third-grade, and fifth-grade assessments: School years 1998–99, 1999–2000, 2001–02, and 2003–04

Proficiency	Reading			Mathematics		
	a	b	c	a	b	c
Level 1	3.50	-1.46	0.0	3.55	-1.93	0.0
Level 2	3.22	-0.90	0.0	3.04	-1.19	0.0
Level 3	3.05	-0.61	0.0	4.30	-0.65	0.0
Level 4	4.25	-0.08	0.0	3.61	-0.04	0.0
Level 5	3.00	0.31	0.0	4.40	0.58	0.0
Level 6	3.50	0.77	0.0	5.90	1.03	0.0
Level 7	5.93	1.06	0.0	4.68	1.45	0.0
Level 8	2.45	1.35	0.0	8.32	1.90	0.0
Level 9	6.13	1.87	0.0	4.24	2.43	0.0

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

www.manaraa.com

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 A38 and A39 in appendix A present subgroup differences with respect to mastery of the level that represents the modal "highest level" score within each round.

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

Section 5.1 described the construction of the longitudinal score scales and IRT calibration of parameters. This section will address the issue of the validity of the score scales as measures of student achievement and growth between fall-kindergarten and spring-fifth grade. The validity issue will be examined from several perspectives:

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

5.2.1 Do the Tests Measure the Right Content?

Evidence for the appropriateness of the tests' content can be obtained from two sources: expert judgments and psychometric results. Chapter 2 describes the design of the tests and development of test frameworks (see section 2.1.2). Curriculum experts and teachers provided input with respect to cognitive skills that are both typically taught and developmentally important. Test frameworks in each subject were developed accordingly, and test items in each set of assessments were selected to conform as closely as possible to framework specifications. Field test item pools and proposed final form item selections were reviewed by experts, and content and presentation of items were modified in response to their recommendations.

www.manaraa.com

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

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

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

Evidence that the tests contained items that were of appropriate difficulty for both the individual children taking them, and in the aggregate for the rounds in which they were administered, can

www.manaraa.com

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

Chapter 4 reviews the operating characteristics of the ECLS-K assessment forms, including the percentages of below-chance (floor effect) and near-perfect (ceiling effect) scores (see section 4.3.1 and table 4-4 for reading; section 4.4.1 and table 4-8 for mathematics; and section 4.5.1 and table 4-12 for science). No floor or ceiling effects were found for the reading and mathematics tests in any round, that is, only a negligible number of children scored below-chance or near-perfect scores on the combined routing and second-stage items. The science test had a borderline floor effect, with about 5 percent of children scoring below-chance in fifth grade.

Appendix B shows the match of the ability distribution for each round to the whole set of items in the assessment versions used in the grade. While each child received only the routing test plus one selected second-stage form in each round, the difficulty of the whole set of items administered in each round (routing items plus *all* second-stage forms) should reflect the ability level of the whole sample for that round. For each subject, appendix B lists all items administered in all rounds of the assessments, sorted in ascending order of item difficulty (IRT "b" parameter). The assessment forms in which each item appeared are also noted. The columns for each round of data collection show the mean and standard deviation of theta, the IRT ability estimate. The asterisks in the columns represent the range of abilities two standard deviations below and above the mean, which should include 95 percent of the sample. For example, fall kindergarten (round 1) children in appendix table B-1 have a reading mean of -1.20 and standard deviation of .50 in the IRT metric. That corresponds to an expected range of ability between -2.20 and -0.20 for 95 percent of test takers. The difficulty of items in the K-1 reading assessment forms includes this range. A few easier items are also present, to prevent floor effects for the lowest achievers in fall kindergarten. Since the K-1 assessment forms were used for the first four rounds, fall-kindergarten through spring-first grade, the range of difficulty of items in the K-1 reading forms had to extend to at least two standard deviations above the round 4 mean, or at least b = 1.18. Several K-1 items have difficulty parameters beyond this point, as a precaution against ceiling effects for the highest achievers in spring-first grade. In each subject area, the difficulty range of the test items administered more than spans the range of two standard deviations below and above the theta mean for the round. The evidence in table B3 is consistent with the findings shown in table 4-12 for fifth-grade science: the low level second-stage

www.manaraa.com

test probably should have contained a few more of the easiest items suitable for the lowest achieving students.

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

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

If each test taker had answered *all* of the items in the kindergarten through fifth-grade item pools at *every* round of data collection, it would be possible to measure the cohesiveness of the scale by observing alpha coefficients and item biserials. Of course, it would have been neither reasonable nor practical to administer the whole item pools to everyone at every round. The IRT "a" parameters provide the same type of insight into the cohesiveness of a set of test items (see section 3.2.1). This parameter represents item discrimination, or the ability of an item to discriminate, or separate, people whose ability level is above or below the calibrated difficulty of the item. In other words, the "a" parameters indicate how strongly each item is related to the underlying construct being measured by the test, with values of 1.0 or above indicating a strong relationship. Values above 1.0 for most of the items in a test constitute evidence that there is a strong underlying factor.

Of the 186 items in the reading scale, only 14 have "a" parameter values less than 1.0, and half of those are picture-vocabulary items. The rest are based on either listening comprehension, understanding conventions of print, or difficult vocabulary words. *All* of the items tapping reading skills, from simple letter recognition and decoding in kindergarten to comprehension of complex reading passages in the later rounds, have "a" parameters above 1.0. Results for mathematics were quite similar, with only six of 153 items having "a" parameters below 1.0. Of these, four were geometry items, which were identified in the field test as being slightly weaker than the other mathematics categories with respect to cohesiveness of the scale, but were included in the item pool to conform to framework specifications. Examination of the reading and mathematics "a" parameters provide evidence that the item

www.manaraa.com

pools and resulting score scales are strongly related to an underlying construct that spans the kindergarten through fifth grade years.

Results for the science assessment are strikingly different, with "a" parameters for nearly three-quarters of the items (68 out of a total of 92) falling below 1.0. This is a consequence of the composition of the science item pool, which is a mix of life science, earth science, and physical science topics. Furthermore, the science assessments did not assume a hierarchical structure in the science curriculum comparable to the patterns for reading and mathematics. In other words, it would be possible for children in some schools to master difficult material relating to the life sciences without having been exposed to basic concepts in earth science, or vice versa. That is the reason that proficiency levels within the science assessments were neither hypothesized nor identified. The relatively low "a" parameters for the science items do not necessarily, however, make IRT methodology inappropriate for calibration of the science scale. In fact, for all except 14 of the 92 items, "a" parameter values were .60 or above. This suggests that although there may be multiple factors influencing item responses, they are all related to each other.

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

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

www.manaraa.com

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

Table 5-7 shows correlations of test scores in each round with scores in the same subject in other rounds. Note that for both reading and mathematics, correlations are highest near the diagonal, and get progressively lower toward the lower left corner of each set. In other words, scores in each subject appear to be most closely related to the most recent or subsequent score, and least closely related to rounds that are more distant.

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

Subject	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Reading						
Round 1	1.00					
Round 2	0.78	1.00				
Round 3	0.77	0.88	1.00			
Round 4	0.66	0.78	0.82	1.00		
Round 5	0.61	0.68	0.70	0.76	1.00	
Round 6	0.58	0.64	0.65	0.73	0.85	1.00
Mathematics						
Round 1	1.00					
Round 2	0.84	1.00				
Round 3	0.81	0.85	1.00			
Round 4	0.73	0.79	0.82	1.00		
Round 5	0.73	0.76	0.78	0.79	1.00	
Round 6	0.70	0.72	0.76	0.78	0.88	1.00
Science						
Round 5	†	†	†	†	1.00	
Round 6	†	†	†	†	0.85	1.00

† Not applicable.
NOTE: Table estimates are based on C1_6SCO panel weight. Science was not tested in kindergarten/first grade.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

For example, the highest correlation (i.e., best predictor) for round 6 reading is the round 5 reading measure, with a correlation coefficient of .85. Previous reading scores are also strongly correlated with round 6 reading, but the relationship becomes weaker going back in time. While reading ability at kindergarten entry is a good predictor of fifth-grade achievement (correlation = .58), other factors present in the intervening years presumably have an important influence as well. Measures of family and school circumstances that relate to student achievement are provided in the ECLS-K database. Exploration of the role these variables play in predicting later achievement is beyond the scope of this report.

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

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

Round	Reading x Mathematics	Reading x Science	Mathematics x Science
Round 1	0.77	†	†
Round 2	0.77	†	†
Round 3	0.75	†	†
Round 4	0.74	†	†
Round 5	0.75	0.72	0.73
Round 6	0.75	0.70	0.75

† Not applicable.
NOTE: Table estimates are based on C1_6SCO panel weight. Science was not tested in kindergarten/first grade.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

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

5.2.5 Comparison of ECLS-K Results With Findings From Other Studies

An additional way to validate the ECLS-K measures would be to compare ECLS-K results with findings of similar studies. Ideally, these "similar studies" would have tests that measure the same content as the ECLS-K tests, and have similar formats, administration procedures, reliabilities, and

www.manaraa.com

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

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

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

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

www.manaraa.com

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

A consensus finding of several studies was that Black-White gaps tend to widen after children enter school (Grissmer et al. 1998; Ferguson 1998). This was consistent with ECLS-K results. In the ECLS-K, the Black-White reading score gap increased only slightly, from .40 to .42 of a standard deviation (SD), by spring-kindergarten, but then expanded to .52 SD by spring first grade, and .71 SD in rounds 5 and 6, when most children were in third- and fifth-grade. A similar pattern was found for mathematics, with an initial fall kindergarten gap of .61 standard deviations widening to .82 and then .85 SD in rounds 5 and 6.

The study that is perhaps most comparable with ECLS-K may be the National Assessment of Educational Progress (NAEP) 2003 assessments in reading and mathematics. Both were large-scale samples representing a national population, in about the same year and similar grades. The content specifications for the ECLS-K tests were derived from NAEP frameworks. Similar IRT methodology was used in producing score scales. Table 5-9 shows reading and mathematics score gaps for selected subgroups for the NAEP 2003 fourth-grade assessment and for ECLS-K rounds 5 and 6, which consisted primarily of third- and fifth-graders. NAEP subgroup differences in reading and mathematics scores were quite similar to the differences found in both ECLS-K rounds for the male/female comparison and for White students compared with Black students. For all of these contrasts except for the male/female difference in mathematics scores the ECLS-K results showed slightly smaller gaps than those found for NAEP fourth-graders. Statistics for White/Hispanic score gaps are included in table 5-9 although race/ethnicity for Hispanic students is defined differently in NAEP and ECLS-K.

www.manaraa.com

Table 5-9. Subgroup gaps in standard deviation units, NAEP and ECLS-K: School years 2001–02, 2002–2003, and 2003–04

Subgroup gaps in standard deviation units	NAEP	ECLS Round 5	ECLS Round 6
Reading			
Female - male	.20	.18	.14
White - Black	.82	.71	.71
White - Hispanic	.76	†	†
White - Hispanic, race specified	†	.49	.43
White - Hispanic, race not specified	†	.78	.72
Mathematics			
Female - male	-.10	-.15	-.16
White - Black	.96	.82	.85
White - Hispanic	.76	†	†
White - Hispanic, race specified	†	.49	.41
White - Hispanic, race not specified	†	.67	.55

† Not applicable.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Reading and Mathematics Assessments, and Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2002 and spring 2004.

As similar as the NAEP and ECLS-K assessments are in many respects, there are also some important differences that relate to the comparability of measurements of gaps:

- NAEP used a cross-sectional sample of children in fourth grade in 2003; the ECLS-K sample was a longitudinal followup of a kindergarten sample. Most of the children tested in round 5 in 2002 were in third grade, and in round 6 in fifth grade, but about 9 percent of children in each round were not yet in the modal grade.
- The NAEP cross sectional sample could be expected to contain more recent immigrants than the ECLS-K longitudinal sample. ECLS-K round 6 children tested in spring 2004 had all joined the sample in kindergarten or first grade, during the 1998-99 or 1999-2000 school year, so they had been attending school in the U.S. for at least four years. The NAEP sample consisted of children in fourth grade in 2003, and included children whose early schooling may have taken place in another country with instruction in a language other than English.
- Tests in ECLS-K were individually administered, while NAEP used group administration.
- NAEP had two different sources for race variables: school records and student self-report (Table 5-9 shows race/ethnicity from school records). ECLS-K used a composite race/ethnicity variable, derived from parent interviews in most cases, and from a variety of other sources when parent reports were unavailable.

www.manaraa.com

- NAEP reported scores for Hispanics as a group, while ECLS-K had separate categories for Hispanic, race specified and Hispanic, race not specified.

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, socioeconomic status (SES), and school type for all of the fifth-grade direct cognitive measures. For measures that can be compared with the analogous scores in earlier rounds, results for rounds 1 through 5 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

www.manaraa.com

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

The standardized scores (T-scores) are also overall measures of status at a point in time, but they are norm-referenced rather than criterion-referenced. They do not answer the question, "What skills do children have?" but rather, "How do they compare with their peers?" The transformation to a familiar metric with a mean of 50 and standard deviation of 10 facilitates comparisons in standard deviation units. T-score means may be used longitudinally to illustrate the increase or decrease in gaps in achievement among subgroups over time. T-scores are not recommended for measuring individual gains over time. The IRT scale scores or proficiency probability scores 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 fifth 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 third and fifth grade at levels 6 or 7, literal inference and extrapolation. The proficiency level at which the largest change is taking place is likely to be different for children with different initial status, background, and school setting. Changes in proficiency probabilities over time may be used to identify the process variables that are effective in promoting achievement gains in specific skills.

www.manaraa.com

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 fifth-grade reading, math, and science assessments are based on 25, 18, and 21 items respectively (15, 17, and 15 items for the same subjects in grade 3; and 20, 16, and 12 items for the K-1 reading, math, and general knowledge assessments). They target specific sets of skills and cover a broad range of difficulty. These scores may be of interest to researchers because they are based on a specific set of assessment items, which was the same for all children who took the fifth-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- and fifth-grade routing tests were each based on different and more difficult sets of items. The fifth-grade routing test number-right scores should *not* be compared with the routing test number-right scores for earlier rounds.

Item cluster scores in reading (e.g., Decoding Score Gr 5) and science (e.g., Life Science Gr 5) are based on a count of the number correct for a particular set of items. Users may wish to relate these scores to process variables to get a perspective that is somewhat different from that of the hierarchical levels of skills. However, with only three to seven items in each of these item cluster scores, reliabilities tend to be relatively low (see sections 4.3.2, 4.3.3, 4.5.2, and 4.5.3).

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

www.manaraa.com

5.3.1.3 Choosing the Correct Sample Weight

The ECLS-K database contains several versions of sample weights, designed to identify students participating in selected rounds and produce national estimates accordingly. Cross sectional weights should be used only when analyzing data from a single round of data collection. When multiple rounds are involved, as in predicting outcomes in later rounds from variables measured earlier, a panel weight is appropriate. Panel weights are defined for specific combinations of rounds. If analysis of round 6 outcomes depends on inputs from *all five* previous rounds, the C1_6SCO panel weight can be selected. This panel weight has a value of zero for any child who did not participate in one or more rounds. It is important to remember that the round 3 (fall-first grade) data collection was based on a small subsample of approximately 30 percent of the longitudinal sample. Selecting the C1_6SCO panel weight will, in effect, delete all cases from the analysis who were not part of the fall-first grade subsample. While weighted estimates may not be affected very much, significance tests depend on *un*weighted sample sizes, so findings of statistical significance, especially for analysis of population subgroups, could be severely affected. If fall-first grade variables are not specifically required, using the C1_6FCO panel weight, which depends on participation in rounds 1, 2, 4, 5, and 6, but not round 3, would increase sample sizes substantially. Additional details on selection and application of sample weights can be found in the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Combined User's Manual for the ECLS-K Fifth-Grade Data Files and Electronic Codebooks* (NCES 2006–032) (Tourangeau et al. forthcoming).

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 11 points and gains of about 21 points between spring-third grade and spring-fifth 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

www.manaraa.com

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 21 points between third grade and fifth grade, again starting from very different initial status. (The similarity in scale score gains between reading and mathematics is coincidental; there is no claim that the same score or amount of gain in different subjects represents a comparable level of achievement or gain.)

It would be incorrect to conclude that because different subgroups of children are gaining quantitatively the same number of scale score points, they are learning the same things, or that these gains are qualitatively comparable in any sense. The problem is nonequivalence of scale units: children who gain 10 or 11 points at the low end of the scale during kindergarten, for example, by mastering letter recognition and letter sounds, are not learning the same things as more advanced children in the same grade, who are achieving their 10-point gains by learning to read words and sentences. Nor can gains in comprehension of reading passages in the later rounds be considered equivalent to gains of the same number of points in basic skills in the early elementary years.

The use of adaptive assessments increases the reliability of individual assessment scores by removing the sources of floor and ceiling effects. When assessment forms are matched to children's ability levels, all 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 nine reading and nine mathematics proficiency levels that were described in chapter 4, model critical stages in the development of skills. Criterion-referenced points serve two purposes at the individual level: (1) they provide information about changes in each child's mastery or proficiency at *each* level, and (2) they provide information about *where* on the scale the child's gain is taking place. This provides analysts with two options for analyzing

www.manaraa.com

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 A33). While most children were making their largest gains between first and third grade at level 5, a small number of children were advancing their skills in solving word problems based on rate and measurement, level 7. The mastery rate for level 7 advanced from near zero at the end of first grade to 13 percent at the end of third grade (shown in table A35). The table breakdowns demonstrate that these proportions and the average gains in the proportions for this particular skill are quite different for subgroups of children defined by various demographic and school-process categories. Similarly, gains at each level between any selected round and a subsequent round may be computed for individual children and treated as outcome variables in multivariate models that include background and process measures.

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

Each different analytical approach provides a different perspective with respect to understanding 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 *ECLS-K Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05).

www.manaraa.com

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 effect on analysis results for long time intervals, such as measuring gains from spring-first grade to spring-third grade, but may be more important within grade, especially fall-to-spring kindergarten. In designing an analysis plan, it is important to consider whether and how differences in ages, assessment dates, and intervals may affect the results, to look at relationships between these factors and other variables of interest, and to compensate for differences if necessary. More details can be found in the *ECLS-K Psychometric Report for Kindergarten Through First Grade* (NCES 2002–05).

www.manaraa.com

This page is intentionally left blank.

www.manaraa.com

6. PSYCHOMETRIC CHARACTERISTICS OF THE SELF-DESCRIPTION QUESTIONNAIRE

Chapter 6 describes the selection and development of the Self-Description Questionnaire (SDQ), which asked children to rate their competence and interest in school subjects and relationships with peers as well as behaviors that might interfere with their academic and social competence. This chapter provides details of the psychometric characteristics of this instrument.

6.1 Self-Description Questionnaire (SDQ)

Beginning in the third-grade data collection in the ECLS-K, children were asked to provide self-assessments of their academic and social skills. In the SDQ, fifth-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." Six scales were produced from the SDQ items. The scale scores on 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 for scores is lower for scales with only six items, and for the Internalizing Problem Behaviors (see table 6-1). This is consistent with other research because internalizing problems are less visible and more difficult to rate. Weighted means and standard deviations for these scales are shown in table 6-2. The mean score in each academic area (Perceived Interest/Competence in Reading, Math, All Subjects) was lower in the fifth-grade data collection than in third grade, as were the mean scores for the Externalizing and Internalizing Problems scales. The mean score for Peer Relations increased slightly from third to fifth grade.

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

www.manaraa.com

Table 6-1. Reliability estimates for scores of the Self-Description Questionnaire (SDQ) scale , spring-fifth grade: School year 2003–04

Description	Number of items	Alpha coefficient
Perceived Interest/Competence — Reading	8	.90
Perceived Interest/Competence — Math	8	.92
Perceived Interest/Competence — All Subjects	6	.83
Perceived Interest/Competence — Peer Relations	6	.82
Externalizing Problems	6	.78
Internalizing Problems	8	.79

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

Table 6-2. Self-Description Questionnaire (SDQ) weighted means and standard deviations, spring-fifth grade: School year 2003–04

Description	Weighted mean	Standard deviation
Perceived Interest/Competence — Reading	3.00	.74
Perceived Interest/Competence — Math	2.92	.78
Perceived Interest/Competence — All Subjects	2.71	.65
Perceived Interest/Competence — Peer Relations	2.98	.63
Externalizing Problems	1.89	.69
Internalizing Problems	2.08	.64

NOTE: Table estimates based on C6CW0 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 2004.

SDQ score statistics for subpopulations are presented in tables 6-3 through 6-8. Children who had been retained (third- or fourth-graders in this round) rated themselves lower in the academic interest/competence areas and rated themselves as having more behavior problems, both internalizing and externalizing problems. Their mean self-rating on peer competence was similar to that of their fifth-grade peers.

www.manaraa.com

Table 6-3. Score breakdown, Self-Description Questionnaire (SDQ), perceived interest/competence in reading, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,148	3.00	0.74	1,077	2.96	0.74
Sex						
Male	5,007	2.91	0.75	645	2.87	0.76
Female	5,141	3.10	0.72	432	3.10	0.68
Race/ethnicity						
White, non-Hispanic	5,974	3.01	0.74	473	2.94	0.75
Black, non-Hispanic	1,012	3.02	0.76	257	3.02	0.75
Hispanic, race specified	914	2.98	0.74	106	2.88	0.65
Hispanic, race not specified	960	2.93	0.73	115	3.02	0.70
Asian	729	3.05	0.70	48	2.73	0.66
Hawaiian, other Pacific Islander	135	2.75	0.70	9	2.99	0.62
American Indian/Alaska Native	158	3.07	0.83	50	2.94	0.70
More than one race, non-Hispanic	252	3.02	0.68	17	2.87	0.70
Socioeconomic status						
First quintile (lowest)	1,341	2.93	0.72	364	3.01	0.74
Second quintile	1,668	2.93	0.77	243	2.87	0.80
Third quintile	1,835	2.98	0.78	149	2.87	0.69
Fourth quintile	2,176	3.00	0.72	121	3.01	0.70
Fifth quintile (highest)	2,431	3.16	0.71	88	3.05	0.63
School type						
Public school	8,182	3.00	0.74	981	2.95	0.74
Private school	1,948	3.02	0.76	93	3.10	0.61

[1] Standard deviation.
NOTE: Table estimates are based on C6CW0 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 2004.

www.manaraa.com

Table 6-4. Score breakdown, Self-Description Questionnaire (SDQ), perceived interest/competence in mathematics, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,148	2.90	0.78	1,077	3.02	0.80
Sex						
Male	5,007	2.96	0.78	645	3.07	0.78
Female	5,141	2.84	0.78	432	2.96	0.83
Race/ethnicity						
White, non-Hispanic	5,974	2.89	0.77	473	2.96	0.80
Black, non-Hispanic	1,012	2.89	0.85	257	3.11	0.85
Hispanic, race specified	914	2.90	0.76	106	3.06	0.63
Hispanic, race not specified	960	2.95	0.78	115	3.07	0.82
Asian	729	3.00	0.70	48	2.63	0.83
Hawaiian, other Pacific Islander	135	2.69	0.77	9	2.83	0.70
American Indian/Alaska Native	158	2.75	0.74	50	3.13	0.72
More than one race, non-Hispanic	252	2.96	0.79	17	3.07	0.58
Socioeconomic status						
First quintile (lowest)	1,341	2.93	0.79	364	2.95	0.83
Second quintile	1,668	2.85	0.82	243	2.95	0.85
Third quintile	1,835	2.85	0.80	149	3.07	0.78
Fourth quintile	2,176	2.91	0.75	121	3.27	0.58
Fifth quintile (highest)	2,431	2.97	0.74	88	3.09	0.77
School type						
Public school	8,182	2.92	0.78	981	3.01	0.80
Private school	1,948	2.79	0.78	93	3.16	0.81

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 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 2004.

www.manaraa.com

Table 6-5. Score breakdown, Self-Description Questionnaire (SDQ), perceived interest/competence in all subjects, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,148	2.71	0.64	1,077	2.71	0.67
Sex						
Male	5,007	2.64	0.65	645	2.63	0.70
Female	5,141	2.78	0.63	432	2.81	0.62
Race/ethnicity						
White, non-Hispanic	5,974	2.71	0.64	473	2.64	0.68
Black, non-Hispanic	1,012	2.69	0.70	257	2.80	0.69
Hispanic, race specified	914	2.68	0.60	106	2.70	0.56
Hispanic, race not specified	960	2.76	0.60	115	2.81	0.64
Asian	729	2.84	0.66	48	2.37	0.57
Hawaiian, other Pacific Islander	135	2.47	0.58	9	2.52	0.58
American Indian/Alaska Native	158	2.56	0.64	50	2.84	0.71
More than one race, non-Hispanic	252	2.77	0.68	17	2.72	0.63
Socioeconomic status						
First quintile (lowest)	1,341	2.68	0.66	364	2.70	0.67
Second quintile	1,668	2.68	0.66	243	2.64	0.72
Third quintile	1,835	2.64	0.64	149	2.69	0.70
Fourth quintile	2,176	2.73	0.59	121	2.83	0.60
Fifth quintile (highest)	2,431	2.82	0.64	88	2.70	0.60
School type						
Public school	8,182	2.71	0.64	981	2.70	0.68
Private school	1,948	2.67	0.65	93	2.78	0.53

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 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 2004.

www.manaraa.com

Table 6-6. Score breakdown, Self-Description Questionnaire (SDQ), perceived interest/competence in peer relations, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,148	2.98	0.62	1,077	2.97	0.69
Sex						
Male	5,007	2.93	0.62	645	2.91	0.72
Female	5,141	3.03	0.62	432	3.04	0.65
Race/ethnicity						
White, non-Hispanic	5,974	2.99	0.60	473	2.90	0.69
Black, non-Hispanic	1,012	3.05	0.67	257	3.06	0.74
Hispanic, race specified	914	2.96	0.61	106	2.90	0.66
Hispanic, race not specified	960	2.88	0.64	115	3.14	0.55
Asian	729	2.85	0.59	48	2.64	0.74
Hawaiian, other Pacific Islander	135	2.69	0.62	9	2.89	0.47
American Indian/Alaska Native	158	3.02	0.67	50	2.86	0.67
More than one race, non-Hispanic	252	3.02	0.63	17	3.21	0.54
Socioeconomic status						
First quintile (lowest)	1,341	2.90	0.66	364	2.92	0.69
Second quintile	1,668	2.94	0.66	243	2.97	0.69
Third quintile	1,835	2.97	0.62	149	3.15	0.75
Fourth quintile	2,176	3.03	0.59	121	2.97	0.63
Fifth quintile (highest)	2,431	3.08	0.56	88	3.07	0.63
School type						
Public school	8,182	2.98	0.62	981	2.96	0.70
Private school	1,948	2.99	0.60	93	3.01	0.58

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 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, spring 2004.

www.manaraa.com

Table 6-7. Score breakdown, Self-Description Questionnaire (SDQ), externalizing problems, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,148	1.85	0.68	1077	2.17	0.70
Sex						
Male	5,007	1.99	0.69	645	2.22	0.71
Female	5,141	1.71	0.63	432	2.10	0.69
Race/ethnicity						
White, non-Hispanic	5,974	1.76	0.63	473	2.08	0.70
Black, non-Hispanic	1,012	2.09	0.76	257	2.28	0.70
Hispanic, race specified	914	1.93	0.68	106	2.37	0.70
Hispanic, race not specified	960	1.97	0.70	115	2.08	0.65
Asian	729	1.67	0.55	48	2.04	0.64
Hawaiian, other Pacific Islander	135	2.19	0.67	9	2.45	0.57
American Indian/Alaska Native	158	2.12	0.68	50	2.54	0.59
More than one race, non-Hispanic	252	1.88	0.74	17	2.16	0.93
Socioeconomic status						
First quintile (lowest)	1,341	2.13	0.77	364	2.25	0.72
Second quintile	1,668	1.95	0.73	243	2.21	0.67
Third quintile	1,835	1.87	0.66	149	2.18	0.75
Fourth quintile	2,176	1.75	0.61	121	2.01	0.64
Fifth quintile (highest)	2,431	1.62	0.51	88	1.72	0.61
School type						
Public school	8,182	1.86	0.68	981	2.18	0.71
Private school	1,948	1.76	0.62	93	1.98	0.60

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 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 2004.

www.manaraa.com

Table 6-8. Score breakdown, Self-Description Questionnaire (SDQ), internalizing problems, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	10,148	2.04	0.63	1,077	2.33	0.67
Sex						
Male	5,007	2.02	0.62	645	2.29	0.67
Female	5,141	2.06	0.64	432	2.40	0.65
Race/ethnicity						
White, non-Hispanic	5,974	1.94	0.59	473	2.13	0.62
Black, non-Hispanic	1,012	2.16	0.68	257	2.51	0.64
Hispanic, race specified	914	2.19	0.64	106	2.48	0.69
Hispanic, race not specified	960	2.28	0.65	115	2.65	0.58
Asian	729	2.02	0.59	48	2.46	0.78
Hawaiian, other Pacific Islander	135	2.33	0.62	9	2.23	0.54
American Indian/Alaska Native	158	2.09	0.64	50	2.50	0.66
More than one race, non-Hispanic	252	1.91	0.55	17	2.36	0.79
Socioeconomic status						
First quintile (lowest)	1,341	2.33	0.68	364	2.48	0.66
Second quintile	1,668	2.10	0.65	243	2.36	0.66
Third quintile	1,835	2.02	0.62	149	2.19	0.66
Fourth quintile	2,176	1.91	0.54	121	2.03	0.65
Fifth quintile (highest)	2,431	1.88	0.54	88	2.01	0.54
School type						
Public school	8,182	2.05	0.64	981	2.35	0.67
Private school	1,948	1.97	0.57	93	2.15	0.64

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 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 2004.

Intercorrelations with other scales are presented in table 7-13 in chapter 7.

www.manaraa.com

7. PSYCHOMETRIC CHARACTERISTICS OF THE INDIRECT MEASURES

Chapter 7 describes the selection and development of the fifth-grade indirect measures. The indirect measures were teacher evaluations of children's academic and social skills. This chapter provides details of the psychometric characteristics of these instruments. In addition, the relationships between the direct and indirect cognitive measures are explored.

7.1 Teacher Measures

In the spring-fifth grade data collection (round 6), teachers of the sampled children were asked to evaluate each child's academic and social skills. The fifth-grade teacher measures were similar in design to those administered in kindergarten, first, and third grades, 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 fifth-grade scores, while sharing names with the measures collected earlier, are scaled differently. They should not be directly compared with kindergarten through third-grade scores for the purpose of evaluating gains over time. Data collected in the earlier rounds may, however, be used as covariates in analyzing fifth-grade achievement and behavioral data. Details of the kindergarten, first-grade, and third-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) (Rock and Pollack 2002) and the *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062) (Pollack et al. 2005).

Differential item functioning (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

www.manaraa.com

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. The order of item difficulties was examined by subgroup to check for any major problems in how the items were interpreted in relation to the students who were rated. No problems were identified. The ordering of item difficulties was similar with all subgroups, and the item difficulties clustered closely together.

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.

7.1.1 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. For each of the scales, the child's primary teacher in the area completed the ratings.

www.manaraa.com

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., "reduces fractions to lowest denominator") to others with a more norm-referenced point of view (e.g., "uses various strategies to gain information"). 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 of this report for additional details on the design and development of the ARS instrument.

The Rating Scale model used to estimate ARS scores is described in detail in chapter 3. The reliability of scores for each of the scales is very high (table 7-1). The summary fit statistics for persons and items are acceptable for all the scales (table 7-2). The fit statistics for the step calibrations indicate that the lowest category ("Not yet") was used less than expected.

Table 7-1. Academic Rating Scale (ARS) person reliability for the Rasch-based score, spring-fifth grade: School year 2003–04

Scale	Reliability
Language and Literacy	.95
Mathematical Thinking	.92
Science	.94

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

www.manaraa.com

Table 7-2. Academic Rating Scale (ARS) fit statistics for persons and items, spring-fifth grade: School year 2003–04

Scale	Infit MNSQ[1]	Outfit MNSQ
Persons		
Language and Literacy	.99	1.00
Mathematical Thinking	1.00	.98
Science	.97	.97
Items		
Language and Literacy	1.00	1.00
Mathematical Thinking	1.02	1.00
Science	1.00	.98

[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 2004.

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 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 fifth-grade ARS ratings were different from the items used in the kindergarten, first-grade, and third-grade instruments. Although the fifth-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, "demonstrates an understanding of place value" appeared in the versions of the ARS, first through fifth grade. In spring-first grade, this item was described as "by explaining that fourteen is ten plus four, or using two stacks of ten and five single cubes to represent the number 25" while the fifth-grade ARS described this item stem as "compares decimals to the thousandths place (1.04 > 1.009)." Obviously, a spring-first grade rating with respect to the first description does not represent the same level of skill as the same rating based on the fifth-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 fifth-grade ratings with those from earlier rounds. Covariance

www.manaraa.com

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, the score was set to missing.

The weighted means and standard deviations for the fifth-grade ARS scores are shown in table 7-3. Score breakdowns for population subgroups are presented in tables 7-18 through 7-21 at the end of this chapter.

Table 7-3. Academic Rating Scale (ARS) means and standard deviations, spring-fifth grade: School year 2003–04

Scale	Weighted mean	Standard deviation
Language and Literacy	3.36	.84
Mathematical Thinking	3.37	.70
Science	3.28	.89

NOTE: Table estimates are based on C6CW0 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 2004.

7.1.1.1 Floor and Ceiling

As noted in the section on the development of the ARS, the criteria for some of the items 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 7-4 presents the

www.manaraa.com

percentage of children at the ceiling and floor of the measures. The percentages of perfect scores for literacy and mathematics are somewhat higher than had been found for the same scales in the third-grade ARS, and the percentages of minimum scores in literacy and mathematics are lower than what was found in earlier rounds. The percentages of maximum and minimum scores on science are comparable across the grades. The slight ceiling effect in ARS scores may attenuate correlations with other variables, particularly in analyses focusing on high achieving students.

Table 7-4. Percent of sample with perfect and minimum Academic Rating Scale scores, spring-fifth grade: School year 2003–04

Description	Percent
Perfect scores	
Language and Literacy	6.3
Mathematical Thinking	5.8
Science	6.4
Minimum scores	
Language and Literacy	0.4
Mathematical Thinking	0.4
Science	1.1

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 2004.

Tables 7-5 to 7-7 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 ordering of the items in difficulty is consistent with what would be expected based on reviews of curriculum. The range of difficulty is more limited than expected.

www.manaraa.com

Table 7-5. Academic Rating Scale language and literacy item difficulties (arranged in order of difficulty), spring-fifth grade: School year 2003–04

Item difficulty	Item number and abbreviated content
2.76	Q2. Understands and interprets a story or other text read aloud
2.77	Q4. Reads fluently
2.86	Q1. Conveys ideas clearly when speaking
2.96	Q5. Reads and comprehends expository text
3.04	Q6. Composes multi-paragraph stories/reports with an understandable beginning, middle, and end
3.06	Q3. Uses various strategies to gain information
3.07	Q8. Makes mechanical corrections when reviewing a rough draft
3.22	Q7. 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 2004.

Table 7-6. Academic Rating Scale mathematical thinking item difficulties (arranged in order of difficulty), spring-fifth grade: School year 2003–04

Item difficulty	Item number and abbreviated content
2.27	Q1. Subtracts numbers that require regrouping
2.69	Q6. Shows understanding of place value
2.83	Q9. Divides multi-digit problems with remainders in the quotient
2.99	Q7. Makes reasonable estimates of quantities and checks answers
3.04	Q5. Uses measuring tools accurately
3.04	Q8. Uses strategies to multiply and divide
3.14	Q4. Recognizes properties of shapes such as area, perimeter, and volume
3.15	Q2. Reduces fractions to lowest denominator
3.19	Q10. Demonstrates algebraic thinking
3.29	Q3. Demonstrates money management skills

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

www.manaraa.com

Table 7-7. Academic Rating Scale (ARS) science item difficulties (arranged in order of difficulty), spring-fifth grade: School year 2003–04

Item difficulty	Item number and abbreviated content
2.72	Q3. Classifies and compares living and non-living things in different ways
2.82	Q7. Demonstrates understanding of life science concepts
2.91	Q5. Applies scientific principles to experiences of daily living
2.91	Q1. Makes logical predictions when conducting scientific investigations
2.95	Q4. Forms explanations and conclusions based on observation and investigation
3.03	Q2. Communicates scientific information
3.09	Q6. Demonstrates understanding of physical science concepts

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99, spring 2004.

Tables 7-8 to 7-10 provide standard errors (SE) for each of the ARS scores for fifth grade. The "Score" column is the sum of the raw score ratings. "Measure" is the score estimated using the Rating Scale model. The column labeled "SE" is the corresponding standard error of measurement for those scores. These standard errors can be used in analytic models to correct for the heteroskedasticity of scores.

Table 7-8. Academic Rating Scale language and literacy standard errors, spring-fifth grade: School year 2003–04

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
8	1.00E	.42	19	2.41	.14	30	3.58	.15
9	1.30	.24	20	2.51	.15	31	3.69	.16
10	1.49	.19	21	2.60	.15	32	3.81	.16
11	1.63	.16	22	2.71	.15	33	3.92	.16
12	1.74	.15	23	2.82	.16	34	4.03	.16
13	1.84	.15	24	2.94	.16	35	4.14	.16
14	1.94	.15	25	3.05	.16	36	4.25	.16
15	2.03	.15	26	3.17	.16	37	4.37	.17
16	2.13	.15	27	3.27	.15	38	4.50	.19
17	2.22	.15	28	3.38	.15	39	4.70	.24
18	2.32	.14	29	3.48	.15	40	5.00E	.42

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, spring 2004.

www.manaraa.com

Table 7-9. Academic Rating Scale mathematical thinking standard errors, spring-fifth grade: School year 2003–04

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
10	1.00E	.54	24	2.57	.12	38	3.44	.14
11	1.38	.31	25	2.62	.12	39	3.51	.15
12	1.62	.23	26	2.68	.13	40	3.59	.15
13	1.77	.20	27	2.73	.13	41	3.66	.15
14	1.89	.18	28	2.79	.13	42	3.74	.15
15	1.99	.16	29	2.84	.13	43	3.83	.16
16	2.08	.15	30	2.90	.13	44	3.92	.16
17	2.15	.14	31	2.96	.13	45	4.91	.17
18	2.22	.14	32	3.03	.14	46	4.12	.18
19	2.29	.13	33	3.09	.14	47	4.24	.20
20	2.35	.13	34	3.16	.14	48	4.39	.23
21	2.40	.13	35	3.23	.14	49	4.63	.30
22	2.46	.13	36	3.29	.14	50	5.00E	.53
23	2.51	.12	37	3.36	.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 2004.

Table 7-10. Academic Rating Scale science standard errors: School year 2003–04

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
7	1.00E	.43	17	2.37	.15	27	3.60	.19
8	1.30	.25	18	2.48	.16	28	3.76	.20
9	1.51	.19	19	2.59	.16	29	3.93	.19
10	1.64	.17	20	2.71	.17	30	4.07	.18
11	1.76	.16	21	2.84	.17	31	4.21	.17
12	1.87	.15	22	2.97	.17	32	4.34	.18
13	1.97	.15	23	3.09	.17	33	4.49	.19
14	2.07	.15	24	3.21	.17	34	4.69	.25
15	2.17	.15	25	3.33	.17	35	5.00E	.43
16	2.27	.15	26	3.46	.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 2004.

www.manaraa.com

7.1.2 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, fifth-grade reading 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- and fifth-grade SRS, and two new items were added. The same form was used in third and fifth grades.

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- and fifth-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 for scores on that scale. Thus, this item was included in the Approaches to Learning scale.

In third grade and again in fifth grade, examination of the responses suggested a different perception of a student's self-control and interpersonal social abilities when compared with kindergarten and first grade. As a result, an additional scale was created. The Self-Control scale includes items on control of attention as well as control of emotions and behavior in interactions. The Interpersonal scale included interactions with both adults and peers. In both third and fifth 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 fifth-grade SRS items were the same as items in the kindergarten-first grade (K-1) instrument and the fifth-grade form was identical to the third-grade form, teachers might

www.manaraa.com

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. Exploratory factor analyses were used to provide evidence of the validity of the scales with this sample. The split-half reliabilities for the scores of the teacher SRS scales were high (table 7-11). Reliabilities are nearly identical for fifth-graders in round 6 and for children who were not yet in fifth grade, so the table contains only reliabilities for the whole sample. These reliabilities are also nearly identical to round 5 results.

Table 7-11. Split-half reliability for the teacher Social Rating Scale (SRS) scores, spring-fifth grade: School year 2003–04

Scale	Split-half reliability
Approaches to Learning	.91
Self-Control	.79
Interpersonal	.88
Externalizing Problem Behaviors	.89
Internalizing Problem Behaviors	.77
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 2004.

Weighted means and standard deviations for these scales are shown in table 7-12. About 90 percent of the children whose teachers provided social ratings data were in fifth grade during the round 6 data collection, and about 10 percent were in third or fourth grade. Numbers in the table are for fifth-graders, with scores for children who at round 6 were still in third or fourth grade shown in parentheses. The number of children who had advanced to sixth or seventh grade by round 6 was too small (less than 0.5 percent) to be analyzed separately. SRS score statistics for subpopulations are presented in tables 7-14 through 7-22 at the end of this chapter, with scores for fifth-graders shown separately from those of children in third and fourth grade.

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

www.manaraa.com

.60 to .81 for fifth-graders. Only the Internalizing Problem Behaviors scale had substantially weaker relationships with the other measures, with correlations of .31 to .40. Patterns of correlations for children who were still in third or fourth grade in the fifth-grade round were very similar to patterns for the on-grade-level children, and were also consistent with results in earlier rounds. The correlations between the internalizing problem scale and the other scales were weaker for third- and fourth-graders, ranging from .22 to .32.

Table 7-12. Teacher Social Rating Scale score means and standard deviations, spring-fifth grade: School year 2003–04

Description	Weighted mean	Standard deviation
Approaches to Learning	3.04 (2.67)	.68 (.68)
Self-Control	3.22 (3.02)	.61 (.63)
Interpersonal	3.06 (2.82)	.65 (.69)
Externalizing Problem Behaviors	1.67 (1.92)	.59 (.69)
Internalizing Problem Behaviors	1.65 (1.82)	.55 (.60)
Peer Relations (Self-Control and Interpersonal Combined)	3.13 (2.90)	.60 (.63)

NOTE: Table estimates based on C6CW0 weight. Numbers outside of parentheses represent children in fifth grade at the time of assessment. Numbers within parentheses represent third- and fourth-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 2004.

7.2 Discriminant and Convergent Validity of the Direct and Indirect Measures

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

Correlations among 12 fifth-grade measures were examined for evidence of convergent and discriminant validity. These measures included three teacher ratings of children's achievement (ARS),

www.manaraa.com

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 7-13. The 12 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.	AppLearn	Teacher SRS factor score for Approaches to Learning
5.	SelfCon	Teacher SRS factor score for Self-Control
6.	InterPers	Teacher SRS factor score for Interpersonal
7.	SDQ Read	Child's self-rating of competence in reading
8.	SDQ Math	Child's self-rating of competence in math
9.	SDQ All	Child's self-rating of competence in all subjects
10.	ReadTheta	Direct cognitive test theta (ability) estimate for Reading
11.	MathTheta	Direct cognitive test theta (ability) estimate for Mathematics
12.	SciTheta	Direct cognitive test theta (ability) estimate for Science

Indirect ARS Lit, ARS Math, and ARS Sci measures have counterparts in measures Read Theta, Math Theta, and Science Theta, the direct cognitive assessment scores. It is instructive to compare the discriminant validity within each of the two sets of cognitive measures (the extent to which scores measuring different constructs should be different), as well as the convergent validity across sets (the extent to which scores should be closely related to other measures of the same construct).

www.manaraa.com

Table 7-13. 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-fifth grade: School year 2003–04

	Round 6											
Measures	ARS Lit	ARS Math	ARS Sci	SRS App Learn	SRS Self Con	SRS Inter Pers	SDQ Read	SDQ Math	SDQ All	Read Theta	Math Theta	Sci Theta
ARS Lit.	1.00											
ARS Math	0.68	1.00										
ARS Sci	0.67	[1]	1.00									
SRSAppLearn	0.58	0.44	0.42	1.00								
SRSSelfCon	0.34	0.25	0.23	0.69	1.00							
SRSInterPers	0.41	0.25	0.31	0.72	0.81	1.00						
SDQ Read	0.27	0.17	0.17	0.22	0.14	0.14	1.00					
SDQ Math	0.12	0.18	0.16	0.18	0.05	0.09	0.11	1.00				
SDQ All	0.24	0.21	0.19	0.32	0.18	0.21	0.52	0.55	1.00			
Read Theta	0.63	0.59	0.55	0.35	0.23	0.22	0.30	0.01	0.12	1.00		
Math Theta	0.58	0.65	0.55	0.34	0.20	0.21	0.11	0.22	0.14	0.75	1.00	
Sci Theta	0.50	0.51	0.50	0.25	0.17	0.17	0.17	0.06	0.07	0.77	0.76	1.00

[1] Children were rated by teachers on the ARS mathematics or the ARS Science, but not both. This cell is empty.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

The correlation for the direct cognitive measure of reading with mathematics is .75. The correlation between the direct reading and mathematics scores had a slight but steady decline from the kindergarten through third-grade round, suggesting the possibility of some divergence of the two skills over time. However, the correlations are higher among the direct measures in fifth grade than they were in third grade. The direct cognitive mathematics and science measures were read to the children so that, as much as possible, the reading demands were removed from the content areas. However, the text was available for children to review and thus children who were better readers may have had additional support from the text. Alternatively, children with more limited literacy skills may spend more of their day working on literacy skills with less time available for the content areas. Children with stronger literacy skills might have the opportunity to read more widely in content areas, and the increased exposure to mathematics and science content might increase the development of the concepts and vocabulary needed for success in the content areas.

From kindergarten through the third-grade data collection, the corresponding correlations for ARS were consistently high. In kindergarten through third grade, the same teacher responded to all areas of the ARS (ARS language/literacy, the ARS mathematical thinking, and the ARS science measure). Thus, there was additional method variance in the correlation. In the fifth grade, the teachers who taught

www.manaraa.com

reading, mathematics, and science rated the children on the relevant ARS form; thus, the ARS ratings may have been completed by different teachers. The correlations of the ARS language/literacy with the ARS mathematics scale and with the ARS science scale were lower for this data collection period when compared with previous data collections and when compared with the relationships among the direct measures.

When one examines the cross-correlations from a convergent validity perspective, patterns are similar to those found in third grade. Relationships are stronger within measures than across measures of similar constructs. One would expect that the direct score in each subject area would be more closely related to the indirect measure of the same subject than to measures of the other subjects. This is true for language/literacy (ARS with direct reading) and mathematical thinking (ARS 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 both third and fifth 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 direct science measure showed similar correlations with all three ARS measures in the fifth grade. In the third grade, the correlation of the direct science was greater with the ARS Literacy than with ARS Mathematics or ARS science.

The indirect ARS measures show consistently higher relationships with teacher-rated behavioral scales such as teacher SRS ratings of approaches to learning, interpersonal behavior, and self-control than do the comparable direct cognitive measures (table 7-13). The higher intercorrelations of the SRS with 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. Only the reading teacher completed the SRS ratings. The SRS scales show the strongest relationships with the literacy related scores, although this is true of both the indirect ARS language and literacy scale and the direct cognitive measure in reading. The differences in relationships are smaller than in previous rounds. As was found in previous rounds, among the teacher rated scales of social skills, the SRS approaches to learning scale has the strongest relationship with each of the ARS scales and direct cognitive scores.

Correlations of children's self-ratings with other measures, while still low, are stronger than in the third grade. In third grade, only the self-rating of reading competence with the teacher rating of

www.manaraa.com

language/literacy and the self-rating of competence in all school subjects with the teacher rating of approaches to learning, reached correlations of .20. The slightly stronger correlation in fifth grade suggests an increased awareness of academic performance. Nevertheless, it continues to appear 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 for the indirect measures are presented in tables 7-14 through 7-22. The means for the ARS will be presented first, followed by the SRS.

www.manaraa.com

Table 7-14. Score breakdown, Academic Rating Scale (ARS), language and literacy, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	9,572	3.45	0.81	1,024	2.79	0.79
Sex						
Male	4,721	3.32	0.81	617	2.72	0.77
Female	4,851	3.58	0.79	407	2.89	0.82
Race/ethnicity						
White, non-Hispanic	5,712	3.52	0.82	452	2.93	0.76
Black, non-Hispanic	955	3.30	0.81	246	2.60	0.87
Hispanic, race specified	837	3.42	0.76	99	2.71	0.63
Hispanic, race not specified	867	3.27	0.77	108	2.83	0.73
Asian	681	3.73	0.82	45	2.69	0.87
Hawaiian, other Pacific Islander	122	3.38	0.76	9	3.18	0.72
American Indian/Alaska Native	146	3.16	0.93	46	2.30	0.61
More than one race, non-Hispanic	238	3.44	0.74	17	2.78	0.55
Socioeconomic status						
First quintile (lowest)	1,249	3.13	0.80	346	2.52	0.71
Second quintile	1,568	3.28	0.79	231	2.75	0.76
Third quintile	1,734	3.40	0.78	143	2.86	0.73
Fourth quintile	2,072	3.59	0.78	115	3.09	0.79
Fifth quintile (highest)	2,297	3.81	0.76	83	3.47	0.59
School type						
Public school	7,711	3.44	0.82	933	2.75	0.79
Private school	1,861	3.50	0.79	91	3.34	0.68

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.

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

www.manaraa.com

Table 7-15. Score breakdown, Academic Rating Scale (ARS), mathematical thinking, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	4,735	3.45	0.67	480	2.77	0.63
Sex						
Male	2,313	3.45	0.71	284	2.72	0.64
Female	2,422	3.45	0.63	196	2.83	0.61
Race/ethnicity						
White, non-Hispanic	2,850	3.48	0.68	198	2.90	0.59
Black, non-Hispanic	457	3.32	0.61	107	2.53	0.64
Hispanic, race specified	408	3.52	0.63	54	2.82	0.48
Hispanic, race not specified	431	3.35	0.65	57	2.93	0.67
Asian	331	3.69	0.71	21	2.49	0.48
Hawaiian, other Pacific Islander	67	3.46	0.55	5	3.27	0.71
American Indian/Alaska Native	79	3.09	0.75	27	2.24	0.44
More than one race, non-Hispanic	112	3.59	0.62	11	2.70	0.59
Socioeconomic status						
First quintile (lowest)	611	3.20	0.65	169	2.60	0.70
Second quintile	760	3.34	0.67	101	2.68	0.48
Third quintile	845	3.39	0.58	63	2.83	0.64
Fourth quintile	1,077	3.53	0.64	58	3.04	0.48
Fifth quintile (highest)	1,123	3.76	0.70	38	3.06	0.44
School type						
Public school	3,809	3.44	0.67	447	2.74	0.62
Private school	926	3.50	0.66	33	3.33	0.43

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.

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

www.manaraa.com

Table 7-16. Score breakdown, Academic Rating Scale (ARS), science, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	4,538	3.35	0.88	482	2.82	0.78
Sex						
Male	2,258	3.35	0.90	290	2.84	0.83
Female	2,280	3.36	0.85	192	2.79	0.70
Race/ethnicity						
White, non-Hispanic	2,715	3.48	0.87	232	2.91	0.78
Black, non-Hispanic	468	3.19	0.85	123	2.60	0.74
Hispanic, race specified	397	3.13	0.81	37	2.93	0.79
Hispanic, race not specified	407	3.01	0.92	46	2.85	0.73
Asian	315	3.63	0.86	20	2.88	0.69
Hawaiian, other Pacific Islander	51	3.02	0.76	3	2.29	0.54
American Indian/Alaska Native	69	3.02	0.73	16	2.47	0.80
More than one race, non-Hispanic	116	3.47	0.80	5	3.36	0.97
Socioeconomic status						
First quintile (lowest)	579	2.89	0.83	156	2.61	0.72
Second quintile	761	3.16	0.83	114	2.81	0.76
Third quintile	843	3.36	0.85	68	2.79	0.83
Fourth quintile	934	3.51	0.87	52	3.26	0.97
Fifth quintile (highest)	1,119	3.74	0.80	43	3.23	0.76
School type						
Public school	3,657	3.33	0.88	435	2.80	0.78
Private school	881	3.50	0.81	47	3.21	0.69

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.

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

www.manaraa.com

Table 7-17. Score breakdown, Teacher Social Rating Scale (SRS), approaches to learning, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	9,622	3.04	0.68	1,024	2.67	0.68
Sex						
Male	4,741	2.85	0.68	614	2.56	0.64
Female	4,881	3.24	0.62	410	2.82	0.71
Race/ethnicity						
White, non-Hispanic	5,740	3.09	0.67	451	2.77	0.64
Black, non-Hispanic	957	2.82	0.68	247	2.50	0.73
Hispanic, race specified	852	3.07	0.67	99	2.57	0.56
Hispanic, race not specified	872	3.05	0.68	107	2.73	0.71
Asian	683	3.42	0.56	45	2.85	0.77
Hawaiian, other Pacific Islander	119	2.93	0.79	9	2.91	0.52
American Indian/Alaska Native	147	2.82	0.63	47	2.53	0.64
More than one race, non-Hispanic	239	2.96	0.72	17	2.89	0.62
Socioeconomic status						
First quintile (lowest)	1,250	2.90	0.70	347	2.57	0.69
Second quintile	1,580	2.98	0.68	233	2.71	0.59
Third quintile	1,749	2.96	0.69	141	2.69	0.59
Fourth quintile	2,080	3.11	0.66	115	2.81	0.63
Fifth quintile (highest)	2,303	3.28	0.61	83	3.26	0.60
School type						
Public school	7,729	3.03	0.68	933	2.65	0.68
Private school	1,893	3.14	0.64	91	3.06	0.62

[1] Standard deviation.
NOTE: Table estimates are based on C6CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

www.manaraa.com

Table 7-18. Score breakdown, Teacher Social Rating Scale (SRS), self-control by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	9,533	3.22	0.61	1,013	3.02	0.63
Sex						
Male	4,697	3.08	0.63	604	2.98	0.64
Female	4,836	3.36	0.55	409	3.07	0.62
Race/ethnicity						
White, non-Hispanic	5,698	3.26	0.58	446	3.13	0.58
Black, non-Hispanic	952	3.00	0.68	245	2.84	0.68
Hispanic, race specified	833	3.26	0.58	97	2.88	0.63
Hispanic, race not specified	865	3.24	0.58	105	3.17	0.59
Asian	673	3.50	0.46	45	3.20	0.42
Hawaiian, other Pacific Islander	119	3.09	0.70	9	3.43	0.37
American Indian/Alaska Native	146	3.05	0.55	47	2.60	0.62
More than one race, non-Hispanic	236	3.16	0.67	17	3.20	0.61
Socioeconomic status						
First quintile (lowest)	1,231	3.10	0.64	343	2.91	0.65
Second quintile	1,562	3.19	0.63	231	3.07	0.60
Third quintile	1,732	3.16	0.61	140	3.07	0.60
Fourth quintile	2,066	3.24	0.58	113	3.18	0.59
Fifth quintile (highest)	2,288	3.40	0.53	82	3.30	0.57
School type						
Public school	7,660	3.21	0.61	923	3.01	0.63
Private school	1,873	3.29	0.55	90	3.12	0.71

[1] Standard deviation.
NOTE: Table estimates are based on C6CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

www.manaraa.com

Table 7-19. Score breakdown, Teacher Social Rating Scale (SRS), interpersonal, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	9,439	3.06	0.65	986	2.82	0.69
Sex						
Male	4,627	2.88	0.65	590	2.76	0.66
Female	4,812	3.23	0.60	396	2.90	0.71
Race/ethnicity						
White, non-Hispanic	5,665	3.08	0.65	440	2.93	0.65
Black, non-Hispanic	932	2.90	0.68	239	2.61	0.73
Hispanic, race specified	817	3.14	0.59	93	2.80	0.62
Hispanic, race not specified	846	3.09	0.61	101	2.93	0.69
Asian	673	3.33	0.56	44	2.96	0.53
Hawaiian, other Pacific Islander	114	2.94	0.66	9	3.29	0.44
American Indian/Alaska Native	145	2.81	0.65	41	2.51	0.66
More than one race, non-Hispanic	235	3.04	0.63	17	2.90	0.57
Socioeconomic status						
First quintile (lowest)	1,207	2.95	0.66	331	2.76	0.69
Second quintile	1,543	3.02	0.64	225	2.84	0.66
Third quintile	1,711	2.99	0.67	137	2.81	0.60
Fourth quintile	2,052	3.09	0.63	110	2.98	0.64
Fifth quintile (highest)	2,279	3.24	0.60	81	3.24	0.52
School type						
Public school	7,570	3.04	0.65	896	2.81	0.68
Private school	1,869	3.16	0.60	90	2.98	0.72

[1] Standard deviation.
NOTE: Table estimates are based on C6CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

www.manaraa.com

Table 7-20. Score breakdown, Teacher Social Rating Scale (SRS), externalizing problem behaviors, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	9,567	1.67	0.59	1,018	1.92	0.69
Sex						
Male	4,711	1.82	0.63	610	1.96	0.69
Female	4,856	1.51	0.51	408	1.86	0.69
Race/ethnicity						
White, non-Hispanic	5,714	1.64	0.56	450	1.85	0.65
Black, non-Hispanic	949	1.91	0.65	245	2.12	0.78
Hispanic, race specified	842	1.61	0.58	99	1.91	0.57
Hispanic, race not specified	868	1.62	0.59	105	1.73	0.62
Asian	679	1.39	0.44	45	1.62	0.52
Hawaiian, other Pacific Islander	119	1.76	0.70	9	1.86	0.39
American Indian/Alaska Native	145	1.66	0.42	47	2.19	0.68
More than one race, non-Hispanic	238	1.75	0.64	17	1.60	0.48
Socioeconomic status						
First quintile (lowest)	1,237	1.75	0.64	345	1.98	0.71
Second quintile	1,565	1.70	0.60	231	1.84	0.74
Third quintile	1,740	1.74	0.60	141	1.82	0.66
Fourth quintile	2,071	1.63	0.55	113	1.79	0.58
Fifth quintile (highest)	2,298	1.52	0.49	83	1.65	0.46
				345	1.98	0.71
School type						
Public school	7,688	1.68	0.60	927	1.93	0.69
Private school	1,879	1.60	0.52	91	1.74	0.68

[1] Standard deviation.
NOTE: Table estimates are based on C6CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECSL-K), spring 2004.

www.manaraa.com

Table 7-21. Score breakdown, Teacher Social Rating Scale (SRS), internalizing problem behaviors, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	9,475	1.65	0.55	996	1.82	0.60
Sex						
Male	4,656	1.69	0.58	597	1.82	0.63
Female	4,819	1.62	0.52	399	1.82	0.56
Race/ethnicity						
White, non-Hispanic	5,695	1.66	0.56	443	1.87	0.61
Black, non-Hispanic	918	1.68	0.56	241	1.73	0.56
Hispanic, race specified	827	1.64	0.50	96	1.79	0.58
Hispanic, race not specified	863	1.61	0.53	104	1.91	0.70
Asian	666	1.51	0.45	43	1.70	0.48
Hawaiian, other Pacific Islander	116	1.71	0.57	9	1.77	0.46
American Indian/Alaska Native	143	1.71	0.48	42	1.87	0.54
More than one race, non-Hispanic	235	1.65	0.57	16	1.86	0.40
Socioeconomic status						
First quintile (lowest)	1,221	1.71	0.57	339	1.84	0.56
Second quintile	1,541	1.66	0.55	224	1.79	0.61
Third quintile	1,732	1.70	0.57	140	1.73	0.61
Fourth quintile	2,054	1.61	0.54	112	1.98	0.80
Fifth quintile (highest)	2,285	1.56	0.49	81	1.80	0.54
School type						
Public school	7,595	1.67	0.56	907	1.83	0.59
Private school	1,880	1.57	0.51	89	1.77	0.69

[1] Standard deviation.
NOTE: Table estimates are based on C6CW0 weight. The range of possible values is 1 to 5. Subgroup counts do not sum to total because demographic variables are missing for some cases.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), spring 2004.

www.manaraa.com

Table 7-22. Score breakdown, Teacher Social Rating Scale (SRS), peer relations: self-control + interpersonal, by fifth-graders, third- and fourth-graders, and population subgroup: School year 2003–04

Characteristic	Fifth-graders			Third- and fourth-graders		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	9,574	3.13	0.60	1,013	2.90	0.63
Sex						
Male	4,715	2.97	0.61	608	2.85	0.62
Female	4,859	3.29	0.55	405	2.98	0.63
Race/ethnicity						
White, non-Hispanic	5,723	3.16	0.59	448	3.01	0.58
Black, non-Hispanic	951	2.94	0.65	244	2.70	0.68
Hispanic, race specified	839	3.19	0.55	97	2.84	0.59
Hispanic, race not specified	867	3.16	0.56	105	3.02	0.59
Asian	680	3.40	0.48	45	3.06	0.45
Hawaiian, other Pacific Islander	118	3.02	0.66	9	3.35	0.31
American Indian/Alaska Native	147	2.91	0.56	46	2.54	0.62
More than one race, non-Hispanic	236	3.09	0.62	17	3.04	0.56
Socioeconomic status						
First quintile (lowest)	1,232	3.01	0.62	344	2.83	0.65
Second quintile	1,573	3.10	0.60	229	2.94	0.59
Third quintile	1,739	3.06	0.61	139	2.93	0.56
Fourth quintile	2,072	3.16	0.57	114	3.06	0.59
Fifth quintile (highest)	2,299	3.31	0.54	82	3.26	0.49
School type						
Public school	7,685	3.12	0.60	923	2.90	0.62
Private school	1,889	3.22	0.55	90	3.04	0.70

[1] Standard deviation.
NOTE: Table estimates are based on C6CW0 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 2004.

www.manaraa.com

This page is intentionally left blank.

www.manaraa.com

REFERENCES

American Association for the Advancement of Science. (1995). *Benchmarks for Science Literacy*. [on-line]. Available: www.project2061.org.

Atkins-Burnett, S., and Meisels, S. J. (2001). *Measures of Socio-emotional Development in Middle Childhood.* (NCES 2001–03). U.S. Department of Education. Washington, DC: National Center for Education Statistics Working Paper.

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.

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.

Ferguson, R. F. (1998). Can Schools Narrow the Black-White Test Score Gap? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

Gresham, F., and Elliot, S. (1990). *Social Skills Rating System*. Circle Pines, MN: American Guidance Services, Inc.

Grissmer, D., Flanagan, A., and Williamson, S. (1998). Why Did the Black-White Score Gap Narrow in the 1970s and 1980s? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap.* Washington, DC: Brookings Institution Press.

Harcourt Brace. (1995). *Science Anytime.* Orlando, FL: Author.

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

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

Kirsch, I.S., et al. (1993). *Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey*. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

www.manaraa.com

Jencks, C. (1998). Racial Bias in Testing. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

Jencks, C., and Phillips, M. (1998).The Black-White Test Score Gap: An Introduction. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap.* Washington, DC: Brookings Institution Press.

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., and Roche, L. (1995). *The effect of gifted and talented programs on academic self-concept: The big fish strikes again*. American Educational Research Journal, *32*(2): 285–319.

Meisels, S.J., and Perry, N.E. (1996). *How Accurate Are Teacher Judgments of Student's Academic Performance?* (NCES 96–08). U.S. Department of Education. Washington, DC: National Center for Education Statistics Working Paper.

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). (1994a). *Reading Framework for the 1994 National Assessment of Educational Progress*. Washington, DC: U.S. Government Printing Office.

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

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

www.manaraa.com

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

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

Phillips, M., Crouse, J., and Ralph, J. (1998). Does the Black-White Test Score Gap Widen After Children Enter School? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap.* Washington, DC: Brookings Institution Press.

Phillips, M., Brooks-Gunn, J., Duncan, G.J., Klebanov, P., and Crane, J. (1998). Family Background, Parenting Practices, and the Black-White Test Score Gap. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

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

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

Rock, D. A. and Stenner, A. J. (2005). Assessment Issues in the Testing of Children at School Entry. *The Future of Children*, *15*(1).

Rock, D.A., et. al. (1985). *Psychometric Analysis of the NLS-72 and the High School and Beyond Test Batteries*. (NCES 85–217). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Rock, D.A., et. al. (1995). *Psychometric Report for the NELS: 88 Base Year Test Battery.* (NCES 95–382). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

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

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

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

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.

www.manaraa.com

Steele, C. M., and Aronson, J. (1998). Stereotype Threat and the Test Performance of Academically Successful African Americans. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap.* Washington, DC: Brookings Institution Press.

Tourangeau, K., Brick, M., Lê, T., Wan, S., Weant, M., Nord, C. Vaden-Kiernan, N., Hagedorn, M., Bissett, E., Dulaney, R., Fowler, J., Pollack, J. Rock, D., Weiss, M., Atkins-Burnett, S., Bose, J., Germino Hausken, E., West, J., Denton, K., Rathbun, A., and Walston, J. (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 Codebook* (NCES 2003–003). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Brick, M., Lê, T., Wan, S., Weant, M., Nord, C. Vaden-Kiernan, N., Hagedorn, M., Bissett, E., Dulaney, R., Fowler, J., Pollack, J. Rock, D. Weiss, M., Atkins-Burnett, S., Germino Hausken, E., West, J., Rathbun, A. and Walston, J. (2004). *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). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Burke, J., Lê., T., Wan, S., Weant, M., Nord, C., Vaden-Kiernan, N. Bissett, E., Dulaney, R., Fields, A., Flores-Cervantes. I., Fowler, J., Pollack, J., Rock, R., Atkins-Burnett, S., Meisels, S., Bose, J., West, J., Denton, K., Rathbun, A., and Walston, J. (2002). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), User's Manual for the ECLS-K First Grade Public-Use Data Files and Electronic Codebook* (NCES 2002–135). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wan, S., Bose, J. and West, J. (2002). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K),User's Manual for the ECLS-K Longitudinal Kindergarten–First Grade Public-Use Data File and Electronic Codebook* (NCES 2002–149). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Lê, T. and Nord, C. (Forthcoming). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Fifth-Grade Methodology Report* (NCES 2006–037). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Pollack, J.M., and Atkins-Burnett, S. (Forthcoming). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Combined User's Manual for the ECLS-K Fifth-Grade Data Files and Electronic Codebooks* (NCES 2006-032). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, *National Assessment of Educational Progress Website Data Tool,* retrieved February 28, 2005 from http://nces.ed.gov/nationsreportcard/naepdata/.

Woodcock, R.W., McGrew, K.S., and Werder, J.K. (1994). *Woodcock-McGrew-Werder Mini-Battery of Achievement.* Itasca, IL: Riverside Publishing.

www.manaraa.com

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.

www.manaraa.com

This page is intentionally left blank.

www.manaraa.com

APPENDIX A

SCORE STATISTICS FOR DIRECT COGNITIVE MEASURES FOR SELECTED SUBGROUPS

Table A1. Reading routing test number right, fifth-grade assessment (range of possible values: 0 to 25): School year 2003–04

Characteristic	Round 6		
	Number	Mean	SD[1]
Total sample	11,250	11.39	5.41
Sex			
Male	5,660	11.09	5.48
Female	5,590	11.71	5.33
Race/ethnicity			
White, Non-Hispanic	6,460	12.59	5.30
Black, Non-Hispanic	1,271	9.03	4.98
Hispanic, race specified	1,021	10.54	4.99
Hispanic, race not specified	1,077	8.91	4.74
Asian	785	12.42	5.37
Hawaiian, Other Pacific Islander	144	10.56	5.37
American Indian, Alaska Native	207	8.45	4.96
More than one race, Non-Hispanic	269	12.73	5.36
Socioeconomic status			
1st quintile (lowest)	1,699	7.72	4.55
2nd quintile	1,912	10.24	4.96
3rd quintile	1,989	11.73	4.99
4th quintile	2,308	13.01	4.81
5th quintile (highest)	2,533	14.74	4.71
School type			
Public school	9,177	11.07	5.35
Private school	2,051	13.84	5.15

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 weight. There is no kindergarten/first-grade or third-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 2004.

www.manaraa.com

Table A2. Mathematics routing test number right, fifth-grade assessment (range of possible values: 0 to 18): School year 2003–04

Characteristic	Round 6		
	Number	Mean	SD[1]
Total sample	11,266	9.64	4.87
Sex			
Male	5,672	10.05	4.86
Female	5,594	9.20	4.84
Race/ethnicity			
White, Non-Hispanic	6,467	10.85	4.60
Black, Non-Hispanic	1,275	6.70	4.23
Hispanic, race specified	1,022	8.72	4.67
Hispanic, race not specified	1,080	8.17	4.67
Asian	785	11.65	4.72
Hawaiian, Other Pacific Islander	144	9.41	4.86
American Indian, Alaska Native	208	6.45	4.75
More than one race, Non-Hispanic	269	10.00	4.87
Socioeconomic status			
1st quintile (lowest)	1,708	6.39	4.56
2nd quintile	1,915	8.49	4.40
3rd quintile	1,991	9.88	4.38
4th quintile	2,308	11.15	4.28
5th quintile (highest)	2,534	12.88	3.88
School type			
Public school	9,193	9.45	4.89
Private school	2,052	11.16	4.32

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 weight. There is no kindergarten/first-grade or third-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 2004.

www.manaraa.com

Table A3. Science routing test number right, fifth-grade assessment (range of possible values: 0 to 21): School year 2003–04

Characteristic	Round 6		
	Number	Mean	SD[1]
Total sample	11,264	13.16	4.23
Sex			
Male	5,671	13.57	4.19
Female	5,593	12.71	4.22
Race/ethnicity			
White, Non-Hispanic	6,467	14.56	3.68
Black, Non-Hispanic	1,272	10.23	4.01
Hispanic, race specified	1,022	12.02	3.98
Hispanic, race not specified	1,080	11.18	4.05
Asian	785	13.64	4.32
Hawaiian, Other Pacific Islander	144	11.56	3.83
American Indian, Alaska Native	209	10.16	4.37
More than one race, Non-Hispanic	269	14.03	3.52
Socioeconomic status			
1st quintile (lowest)	1,706	10.13	4.08
2nd quintile	1,915	12.28	3.93
3rd quintile	1,992	13.54	3.51
4th quintile	2,308	14.46	3.62
5th quintile (highest)	2,534	15.93	3.33
School type			
Public school	9,190	12.95	4.25
Private school	2,052	14.80	3.71

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 weight. There is no kindergarten/first-grade or third-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 2004.

www.manaraa.com

Table A4. Reading IRT scale score, K-5 scale (range of possible values: 0 to 186): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	Nr	Mean	SD	N	Mean	SD
Total sample	17,624	29.03	9.78	18,935	40.11	13.38	5,053	46.85	17.21	16,336	70.23	22.45	14,246	116.15	25.60	11,250	136.73	24.26
Sex																		
Male	8,984	28.44	9.91	9,688	39.10	13.31	2,556	45.47	16.98	8,349	68.38	22.81	7,204	113.97	26.28	5,660	135.11	24.99
Female	8,640	29.65	9.60	9,247	41.20	13.38	2,497	48.30	17.33	7,987	72.20	21.88	7,042	118.45	24.66	5,590	138.46	23.34
Race/ethnicity																		
White, Non-Hispanic	10,433	30.40	9.97	11,073	41.99	13.70	2,935	49.41	17.94	9,435	74.32	22.51	8,082	122.89	24.05	6,460	142.66	22.63
Black, Non-Hispanic	2,854	26.52	7.84	2,968	36.35	11.21	782	42.48	13.80	2,371	62.57	19.99	1,840	104.61	23.42	1,271	125.47	23.26
Hispanic, race specified	1,182	27.03	9.28	1,315	38.30	12.22	322	44.59	14.79	1,233	66.28	20.89	1,252	110.45	25.01	1,021	132.19	23.85
Hispanic, race not specified	1,195	25.38	7.27	1,423	35.70	10.88	377	39.95	12.95	1,335	61.53	19.26	1,314	102.94	24.52	1,077	125.08	23.24
Asian	896	33.47	14.38	1,088	46.20	17.72	257	55.54	24.10	1,042	77.58	24.47	956	120.16	23.65	785	140.89	23.09
Hawaiian, Other Pacific Islander	186	27.73	9.59	202	37.49	11.45	93	40.20	13.46	188	66.63	20.64	171	110.01	22.78	144	133.21	22.44
American Indian, Alaska Native	354	23.53	6.38	344	33.81	9.56	126	34.37	10.27	298	55.57	18.24	232	96.97	24.80	207	120.59	27.82
More than one race, Non-Hispanic	476	29.13	11.38	473	39.95	14.63	152	46.35	15.87	397	71.26	22.77	379	117.25	24.82	269	141.69	21.45
Socioeconomic status																		
1st quintile (lowest)	2,594	23.88	5.77	2,917	33.34	8.90	753	37.68	11.09	2,363	57.42	17.75	1,964	98.43	23.42	1,699	118.43	23.95
2nd quintile	3,271	26.67	7.50	3,503	37.35	11.34	925	42.62	13.74	2,796	66.21	20.17	2,230	111.00	23.81	1,912	131.54	22.40
3rd quintile	3,470	28.25	7.83	3,686	39.54	11.34	997	47.17	15.77	3,003	70.52	20.19	2437	117.44	22.68	1,989	139.05	20.50
4th quintile	3,650	30.72	9.68	3,909	42.62	12.98	1,019	50.04	16.67	3,173	75.05	21.01	2,688	124.09	22.62	2,308	144.06	20.73
5th quintile (highest)	3,880	34.95	12.57	4,152	47.68	16.54	1159	56.10	20.85	3,642	83.32	23.56	3,158	133.63	20.58	2,533	152.99	17.69
School type																		
Public school	13,736	28.22	9.17	14,578	39.09	12.66	3,809	45.85	16.69	12,998	68.70	21.88	1,1575	114.76	25.69	9,177	135.42	24.30
Private school	3,888	33.36	11.66	4,357	45.71	15.64	1,042	53.85	18.35	3,279	80.38	22.84	2,623	127.01	22.18	2,051	146.94	21.48

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A5. Mathematics IRT scale score, K-5 scale (range of possible values: 0 to 153): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	22.35	8.69	19,647	32.41	11.44	5,226	39.44	13.67	16,641	56.63	16.97	14,349	90.50	21.89	11,266	111.25	22.40
Sex																		
Male	9,479	22.45	9.22	10,041	32.57	12.04	2,644	39.63	14.60	8,506	57.26	18.03	7,277	92.12	22.53	5,672	113.02	22.35
Female	9,156	22.25	8.09	9,606	32.24	10.77	2,582	39.23	12.62	8,135	55.95	15.73	7,072	88.77	21.06	5,594	109.35	22.29
Race/ethnicity																		
White, Non-Hispanic	10,433	24.54	8.95	11,071	35.29	11.56	2,935	42.79	13.74	9,436	60.92	17.02	8,116	96.33	20.49	6,467	116.77	20.68
Black, Non-Hispanic	2,855	19.22	6.46	2,962	27.74	9.21	781	34.33	11.63	2,371	48.33	14.02	1,871	78.31	20.01	1,275	97.82	21.13
Hispanic, race specified	1,588	19.46	7.32	1,624	29.16	10.08	389	36.63	11.85	1,354	53.07	15.86	1,260	85.55	21.30	1,022	107.52	21.56
Hispanic, race not specified	1,800	17.72	6.48	1,834	26.86	9.34	486	32.46	10.87	1,518	49.69	13.47	1,324	81.76	20.34	1,080	104.39	21.83
Asian	897	26.23	10.62	1,088	36.50	13.21	256	44.20	16.67	1,042	59.46	18.13	956	96.24	22.83	785	119.77	22.05
Hawaiian, Other Pacific Islander	187	20.36	7.37	202	29.03	9.75	93	33.50	9.92	188	49.23	12.89	172	84.58	19.09	144	108.80	21.32
American Indian, Alaska Native	354	17.94	6.81	345	27.56	9.48	126	29.75	11.16	298	48.34	13.79	250	77.55	19.44	208	95.67	23.10
More than one race, Non-Hispanic	473	22.26	8.71	472	32.03	10.82	151	38.14	11.95	397	56.86	16.69	380	91.79	21.58	269	113.03	22.03
Socioeconomic status																		
1st quintile (lowest)	3,269	17.11	5.79	3,426	25.71	8.45	895	31.09	10.94	2,572	47.32	13.78	2,001	76.28	19.47	1,708	95.48	22.94
2nd quintile	3,429	20.15	6.90	3,607	30.10	9.85	942	36.25	11.72	2,839	52.92	15.22	2,250	85.75	20.14	1,915	106.13	20.22
3rd quintile	3,546	22.23	7.34	3,721	32.52	10.11	1,001	39.83	11.58	3,017	56.85	15.31	2,452	91.17	19.54	1,991	112.84	19.00
4th quintile	3,676	24.45	8.27	3,921	35.17	10.83	1,023	42.32	12.00	3,178	60.72	15.73	2,693	97.62	19.89	2,308	118.17	18.92
5th quintile (highest)	3,893	28.31	10.23	4,161	39.69	12.67	1,158	48.60	15.11	3,644	67.43	17.31	3,163	105.19	18.68	2,534	126.11	16.69
School type																		
Public school	14,701	21.62	8.28	15,259	31.54	11.07	3,971	38.61	13.50	13,292	55.68	16.79	11,670	89.81	21.99	9,193	110.31	22.60
Private school	3,934	26.57	9.72	4,388	37.37	12.24	1,043	45.83	13.33	3,286	63.25	16.25	2,631	96.47	20.09	2,052	118.60	18.96

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A6. Science IRT scale score, 3-5 scale (range of possible values: 0 to 92): School years 2001–02 and 2003–04

Characteristic	Round 5			Round 6		
	Number	Mean	SD[1]	N	Mean	SD
Total sample	14,339	43.50	14.14	11,264	56.13	14.87
Sex						
Male	7,267	45.00	14.38	5,671	57.80	14.68
Female	7,072	41.91	13.71	5,593	54.34	14.87
Race/ethnicity						
White, Non-Hispanic	8,110	48.65	12.81	6,467	61.23	12.70
Black, Non-Hispanic	1,869	34.19	11.56	1,272	45.76	14.05
Hispanic, race specified	1,259	37.98	13.24	1,022	51.76	14.63
Hispanic, race not specified	1,325	35.00	12.13	1,080	48.86	14.39
Asian	956	43.78	14.56	785	57.20	15.51
Hawaiian, Other Pacific Islander	172	39.28	12.42	144	50.75	13.55
American Indian, Alaska Native	249	36.54	12.45	209	45.05	15.34
More than one race, Non-Hispanic	379	45.54	12.97	269	59.26	11.92
Socioeconomic status						
1st quintile (lowest)	1,999	33.03	11.62	1,706	44.88	14.36
2nd quintile	2,249	40.84	12.57	1,915	52.85	13.61
3rd quintile	2,452	44.55	12.26	1,992	57.91	11.98
4th quintile	2,692	48.55	12.44	2,308	60.93	12.41
5th quintile (highest)	3,162	53.87	12.21	2,534	65.87	11.98
School type						
Public school	11,657	42.84	14.08	9,190	55.39	14.94
Private school	2,633	48.54	13.45	2,052	61.94	13.10

[1] Standard deviation.

NOTE: Table estimates are based on cross-sectional weights within each round (C5CW0, C6CW0). Estimates for third and fifth grade have been put on a common scale to support comparisons. Science was not tested in kindergarten/first grade. 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 and spring 2004.

www.manaraa.com

Table A7. Reading T-scores, standardized within round (range of possible values: 0 to 96): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	50.00	10.00	18,935	50.00	10.00	5,053	50.00	10.00	16,336	50.00	10.00	14,246	50.00	10.00	11,250	50.00	10.00
Sex																		
Male	8,984	49.21	10.03	9,688	49.06	10.16	2,556	49.01	10.17	8,349	49.06	10.42	7,204	49.15	10.35	5,660	49.37	10.25
Female	8,640	50.83	9.90	9,247	51.01	9.73	2,497	51.04	9.71	7,987	51.00	9.43	7,042	50.91	9.53	5,590	50.67	9.69
Race/ethnicity																		
White, Non-Hispanic	10,433	51.74	9.68	11,073	51.69	9.52	2,935	51.78	9.53	9,435	51.87	9.33	8,082	52.60	9.32	6,460	52.44	9.68
Black, Non-Hispanic	2,854	47.08	9.18	2,968	46.75	9.83	782	47.25	9.44	2,371	46.48	10.41	1,840	45.57	9.31	1,271	45.38	8.99
Hispanic, race specified	1,182	47.48	9.90	1,315	48.52	9.92	322	48.80	9.27	1,233	48.36	9.84	1,252	47.82	9.77	1,021	48.09	9.46
Hispanic, race not specified	1,195	45.54	9.18	1,423	46.16	9.90	377	45.17	9.92	1,335	46.18	9.76	1,314	44.87	9.82	1,077	45.23	8.95
Asian	896	54.03	11.78	1,088	54.10	10.67	257	54.07	11.67	1,042	52.84	10.20	956	51.60	8.99	785	51.58	9.57
Hawaiian, Other Pacific Islander	186	48.15	10.68	202	47.77	9.99	93	45.34	9.33	188	48.75	8.91	171	47.73	8.78	144	48.51	9.03
American Indian, Alaska Native	354	42.90	8.85	344	44.41	9.57	126	40.26	9.79	298	42.94	10.02	232	42.47	10.31	207	43.73	10.77
More than one race, Non-Hispanic	476	49.72	10.69	473	49.65	10.17	152	49.83	9.85	397	50.46	10.08	379	50.46	9.67	269	51.90	9.03
Socioeconomic status																		
1st quintile (lowest)	2,594	43.69	7.86	2,917	44.09	8.98	753	43.52	9.28	2,363	44.02	10.10	1,964	43.12	9.63	1,699	42.70	9.08
2nd quintile	3,271	47.43	8.64	3,503	47.83	9.50	925	47.45	9.27	2,796	48.41	9.66	2,230	48.04	9.31	1,912	47.70	8.78
3rd quintile	3,470	49.48	8.85	3,686	49.95	9.04	997	50.62	9.04	3,003	50.52	8.80	2,437	50.51	8.63	1,989	50.68	8.44
4th quintile	3,650	52.24	9.35	3,909	52.42	8.92	1,019	52.52	8.63	3,173	52.42	8.39	2,688	53.07	8.67	2,308	52.86	8.88
5th quintile (highest)	3,880	56.36	10.25	4,152	55.69	9.48	1,159	55.48	9.32	3,642	55.40	8.44	3,158	56.74	8.10	2,533	56.98	8.55
School type																		
Public school	13,736	49.11	9.73	14,578	49.21	9.85	3,809	49.38	9.91	12,998	49.37	9.99	11,575	49.46	10.05	9,177	49.44	9.92
Private school	3,888	54.82	10.08	4,357	54.31	9.73	1,042	54.62	8.69	3,279	54.30	8.68	2,623	54.21	8.53	2,051	54.34	9.58

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A8. Mathematics T-scores, standardized within round (range of possible values: 0 to 96): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	50.00	10.00	19,647	50.00	10.00	5,226	50.00	10.00	16,641	50.00	10.00	14,349	50.00	10.00	11,266	50.00	10.00
Sex																		
Male	9,479	49.96	10.42	10,041	50.02	10.32	2,644	49.94	10.57	8,506	50.18	10.51	7,277	50.74	10.36	5,672	50.86	10.18
Female	9,156	50.04	9.54	9,606	49.98	9.65	2,582	50.06	9.36	8,135	49.81	9.42	7,072	49.21	9.54	5,594	49.08	9.71
Race/ethnicity																		
White, Non-Hispanic	10,433	52.73	9.49	11,071	52.65	9.37	2,935	52.56	9.21	9,436	52.45	9.38	8,116	52.62	9.36	6,467	52.47	9.59
Black, Non-Hispanic	2,855	46.36	8.66	2,962	45.79	9.21	781	46.16	9.76	2,371	45.13	9.96	1,871	44.50	9.24	1,275	44.01	8.52
Hispanic, race specified	1,588	46.39	9.53	1,624	47.06	9.70	389	48.07	9.55	1,354	48.00	9.91	1,260	47.79	9.71	1,022	48.20	9.32
Hispanic, race not specified	1,800	43.98	9.20	1,834	44.74	9.64	486	44.60	9.72	1,518	46.23	8.98	1,324	46.05	9.32	1,080	46.84	9.13
Asian	897	54.24	10.32	1,088	53.42	9.88	256	53.01	10.45	1,042	51.53	9.83	956	52.66	10.43	785	54.27	10.69
Hawaiian, Other Pacific Islander	187	47.74	9.21	202	47.06	9.21	93	45.88	8.12	188	46.05	8.42	172	47.27	8.77	144	48.83	9.20
American Indian, Alaska Native	354	44.20	9.54	345	45.53	9.46	126	41.86	10.58	298	45.25	9.35	250	44.21	8.83	208	43.40	9.42
More than one race, Non-Hispanic	473	50.04	9.43	472	49.86	9.29	151	49.31	9.16	397	50.12	10.00	380	50.52	9.86	269	50.85	10.10
Socioeconomic status																		
1st quintile (lowest)	3,269	43.25	8.59	3,426	43.68	9.05	895	43.34	9.75	2,572	44.49	9.83	2,001	43.54	9.11	1,708	43.23	9.35
2nd quintile	3,429	47.56	8.94	3,607	48.09	9.35	942	47.81	9.33	2,839	48.00	9.79	2,250	47.86	9.08	1,915	47.43	8.53
3rd quintile	3,546	50.29	8.67	3,721	50.44	8.86	1,001	50.73	8.63	3,017	50.42	8.97	2,452	50.28	8.74	1,991	50.38	8.44
4th quintile	3,676	52.82	8.89	3,921	52.71	8.84	1,023	52.55	8.18	3,178	52.55	8.52	2,693	53.21	9.08	2,308	52.94	8.82
5th quintile (highest)	3,893	56.67	9.43	4,161	56.12	9.15	1,158	56.27	8.86	3,644	55.82	8.36	3,163	56.71	8.71	2,534	57.05	8.64
School type																		
Public school	1,4701	49.15	9.84	15,259	49.24	9.91	3,971	49.37	10.07	13,292	49.46	10.06	11,670	49.68	10.06	9,193	49.59	10.04
Private school	3,934	54.90	9.47	4,388	54.31	9.38	1,043	54.82	7.98	3,286	53.86	8.33	2,631	52.74	9.07	2,052	53.21	8.97

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A9. Science T-scores, standardized within round (range of possible values: 0 to 96): School years 2001–02 and 2003–04

Characteristic	Round 5			Round 6		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	14,339	50.00	10.00	11,264	50.00	10.00
Sex						
Male	7,267	51.05	10.06	5,671	51.15	9.98
Female	7,072	48.88	9.81	5,593	48.77	9.88
Race/ethnicity						
White, Non-Hispanic	8,110	53.65	8.68	6,467	53.37	8.72
Black, Non-Hispanic	1,869	43.38	8.89	1,272	43.17	9.23
Hispanic, race specified	1,259	46.08	9.72	1,022	47.12	9.70
Hispanic, race not specified	1,325	43.95	9.22	1,080	45.18	9.51
Asian	956	50.19	10.10	785	50.82	10.48
Hawaiian, Other Pacific Islander	172	46.98	9.35	144	46.36	8.70
American Indian, Alaska Native	249	45.18	9.12	209	42.51	10.54
More than one race, Non-Hispanic	379	51.54	8.96	269	51.98	7.89
Socioeconomic status						
1st quintile (lowest)	1,999	42.47	9.03	1,706	42.53	9.58
2nd quintile	2,249	48.27	8.95	1,915	47.81	8.84
3rd quintile	2,452	50.90	8.35	1,992	51.05	7.87
4th quintile	2,692	53.59	8.41	2,308	53.13	8.41
5th quintile (highest)	3,162	57.15	8.20	2,534	56.68	8.66
School type						
Public school	11,657	49.53	10.00	9,190	49.50	10.02
Private school	2,633	53.57	9.13	2,052	53.95	9.02

[1] Standard deviation.

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

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 and spring 2004.

www.manaraa.com

Table A10. Reading IRT theta score,K-5 scale (range of possible values: -5 to 5): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	-1.20	0.50	18,935	-0.63	0.50	5,053	-0.39	0.51	16,336	0.22	0.48	14,246	0.96	0.36	11,250	1.26	0.36
Sex																		
Male	8,984	-1.24	0.50	9,688	-0.68	0.51	2,556	-0.44	0.52	8,349	0.17	0.50	7,204	0.93	0.37	5,660	1.24	0.37
Female	8,640	-1.16	0.50	9,247	-0.58	0.49	2,497	-0.34	0.50	7,987	0.27	0.45	7,042	1.00	0.34	5,590	1.29	0.35
Race/ethnicity																		
White, Non-Hispanic	10,433	-1.11	0.49	11,073	-0.55	0.47	2,935	-0.30	0.49	9,435	0.31	0.44	8,082	1.06	0.33	6,460	1.35	0.35
Black, Non-Hispanic	2,854	-1.35	0.46	2,968	-0.79	0.49	782	-0.53	0.48	2,371	0.05	0.49	1,840	0.81	0.33	1,271	1.10	0.33
Hispanic, race specified	1,182	-1.33	0.50	1,315	-0.71	0.49	322	-0.45	0.48	1,233	0.14	0.47	1,252	0.89	0.35	1,021	1.19	0.34
Hispanic, race not specified	1,195	-1.42	0.46	1,423	-0.82	0.49	377	-0.64	0.51	1,335	0.04	0.46	1,314	0.78	0.35	1,077	1.09	0.33
Asian	896	-1.00	0.59	1,088	-0.43	0.53	257	-0.18	0.60	1,042	0.35	0.48	956	1.02	0.32	785	1.32	0.35
Hawaiian, Other Pacific Islander	1,6	-1.29	0.54	202	-0.74	0.50	93	-0.63	0.48	188	0.16	0.42	171	0.88	0.31	144	1.21	0.33
American Indian, Alaska Native	354	-1.56	0.44	344	-0.91	0.48	126	-0.89	0.50	298	-0.12	0.48	232	0.69	0.37	207	1.04	0.39
More than one race, Non-Hispanic	476	-1.21	0.54	473	-0.65	0.51	152	-0.40	0.51	397	0.24	0.48	379	0.98	0.35	269	1.33	0.33
Socioeconomic status																		
1st quintile (lowest)	2,594	-1.52	0.39	2,917	-0.93	0.45	753	-0.72	0.48	2,363	-0.07	0.48	1,964	0.72	0.34	1,699	1.00	0.33
2nd quintile	3,271	-1.33	0.43	3,503	-0.74	0.47	925	-0.52	0.48	2,796	0.14	0.46	2,230	0.89	0.33	1,912	1.18	0.32
3rd quintile	3,470	-1.23	0.44	3,686	-0.63	0.45	997	-0.36	0.46	3,003	0.24	0.42	2,437	0.98	0.31	1,989	1.29	0.31
4th quintile	3,650	-1.09	0.47	3,909	-0.51	0.44	1,019	-0.26	0.44	3,173	0.33	0.40	2,688	1.07	0.31	2,308	1.37	0.32
5th quintile (highest)	3,880	-0.88	0.51	4,152	-0.35	0.47	1,159	-0.11	0.48	3,642	0.48	0.40	3,158	1.20	0.29	2,533	1.52	0.31
School type																		
Public school	13,736	-1.24	0.49	14,578	-0.67	0.49	3,809	-0.42	0.51	12,998	0.19	0.47	11,575	0.94	0.36	9,177	1.24	0.36
Private school	3,888	-0.96	0.51	4,357	-0.42	0.49	1,042	-0.15	0.45	3,279	0.42	0.41	2,623	1.11	0.31	2,051	1.42	0.35

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A11. Mathematics IRT theta score,K-5 scale (range of possible values: -5 to 5): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	-1.14	0.50	19,647	-0.62	0.49	5,226	-0.34	0.50	16,641	0.19	0.46	14,349	0.91	0.42	11,266	1.33	0.48
Sex																		
Male	9,479	-1.14	0.52	10,041	-0.62	0.51	2,644	-0.34	0.53	8,506	0.20	0.48	7,277	0.94	0.43	5,672	1.38	0.48
Female	9,156	-1.14	0.48	9,606	-0.62	0.47	2,582	-0.34	0.47	8,135	0.18	0.43	7,072	0.87	0.40	5,594	1.29	0.46
Race/ethnicity																		
White, Non-Hispanic	10,433	-1.00	0.48	11,071	-0.49	0.46	2,935	-0.21	0.46	9,436	0.30	0.43	8,116	1.02	0.39	6,467	1.45	0.46
Black, Non-Hispanic	2,855	-1.32	0.44	2,962	-0.82	0.45	781	-0.53	0.49	2,371	-0.03	0.45	1,871	0.68	0.38	1,275	1.05	0.41
Hispanic, race specified	1,588	-1.32	0.48	1,624	-0.76	0.47	389	-0.44	0.48	1,354	0.10	0.45	1,260	0.82	0.40	1,022	1.25	0.44
Hispanic, race not specified	1,800	-1.44	0.46	1,834	-0.87	0.47	486	-0.61	0.49	1,518	0.02	0.41	1,324	0.74	0.39	1,080	1.18	0.43
Asian	897	-0.93	0.52	1,088	-0.45	0.48	256	-0.19	0.52	1,042	0.26	0.45	956	1.02	0.43	785	1.54	0.51
Hawaiian, Other Pacific Islander	187	-1.25	0.46	202	-0.76	0.45	93	-0.55	0.41	188	0.01	0.38	172	0.79	0.36	144	1.28	0.44
American Indian, Alaska Native	354	-1.43	0.48	345	-0.84	0.46	126	-0.75	0.53	298	-0.02	0.43	250	0.67	0.37	208	1.02	0.45
More than one race, Non-Hispanic	473	-1.14	0.47	472	-0.62	0.46	151	-0.37	0.46	397	0.20	0.45	380	0.93	0.41	269	1.38	0.48
Socioeconomic status																		
1st quintile (lowest)	3,269	-1.48	0.43	3,426	-0.93	0.44	895	-0.67	0.49	2,572	-0.06	0.45	2,001	0.64	0.38	1,708	1.01	0.44
2nd quintile	3,429	-1.26	0.45	3,607	-0.71	0.46	942	-0.45	0.47	2,839	0.10	0.45	2,250	0.82	0.38	1,915	1.21	0.41
3rd quintile	3,546	-1.12	0.44	3,721	-0.59	0.43	1,001	-0.30	0.43	3,017	0.21	0.41	2,452	0.92	0.36	1,991	1.35	0.40
4th quintile	3,676	-1.00	0.45	3,921	-0.48	0.43	1,023	-0.21	0.41	3,178	0.31	0.39	2,693	1.04	0.38	2,308	1.47	0.42
5th quintile (highest)	3,893	-0.80	0.47	4,161	-0.32	0.45	1,158	-0.03	0.44	3,644	0.46	0.38	3,163	1.19	0.36	2,534	1.67	0.41
School type																		
Public school	14,701	-1.18	0.50	15,259	-0.65	0.49	3,971	-0.37	0.50	13,292	0.17	0.46	11,670	0.89	0.42	9,193	1.32	0.48
Private school	3,934	-0.89	0.48	4,388	-0.41	0.46	1,043	-0.10	0.40	3,286	0.37	0.38	2,631	1.02	0.38	2,052	1.49	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A12. Science IRT theta score, 3-5 scale (range of possible values: -5 to 5): School years 2001–02 and 2003–04

Characteristic	Round 5			Round 6		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	14,339	-0.43	0.86	11,264	0.33	0.90
Sex						
Male	7,267	-0.34	0.87	5,671	0.43	0.90
Female	7,072	-0.53	0.85	5,593	0.22	0.89
Race/ethnicity						
White, Non-Hispanic	8,110	-0.12	0.75	6,467	0.63	0.78
Black, Non-Hispanic	1,869	-1.01	0.77	1,272	-0.29	0.83
Hispanic, race specified	1,259	-0.77	0.84	1,022	0.07	0.87
Hispanic, race not specified	1,325	-0.96	0.80	1,080	-0.11	0.86
Asian	956	-0.42	0.87	785	0.40	0.94
Hawaiian, Other Pacific Islander	172	-0.69	0.81	144	0.00	0.78
American Indian, Alaska Native	249	-0.85	0.79	209	-0.35	0.95
More than one race, Non-Hispanic	379	-0.30	0.77	269	0.50	0.71
Socioeconomic status						
1st quintile (lowest)	1,999	-1.08	0.78	1,706	-0.35	0.86
2nd quintile	2,249	-0.58	0.77	1,915	0.13	0.80
3rd quintile	2,452	-0.36	0.72	1,992	0.42	0.71
4th quintile	2,692	-0.12	0.73	2,308	0.61	0.76
5th quintile (highest)	3,162	0.18	0.71	2,534	0.93	0.78
School type						
Public school	11,657	-0.48	0.86	9,190	0.28	0.90
Private school	2,633	-0.13	0.79	2,052	0.68	0.81

[1] Standard deviation.

NOTE: Table estimates are based on cross-sectional weights within each round (C5CW0, C6CW0). Estimates for third and fifth grade have been put on a common scale to support comparisons. Science was not tested in kindergarten/first grade. 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 and spring 2004.

www.manaraa.com

Table A13. Reading decoding score, third- and fifth-grade assessments (range of possible values: 0 to 4): School years 2001–02 and 2003–04

Characteristic	Round 5			Round 6		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	14,228	1.06	1.24	10,525	2.11	1.38
Sex						
Male	7,198	1.03	1.24	5,219	2.08	1.39
Female	7,030	1.10	1.24	5,306	2.14	1.37
Race/ethnicity						
White, Non-Hispanic	8,074	1.22	1.31	6,206	2.38	1.35
Black, Non-Hispanic	1,836	0.63	0.99	1,094	1.53	1.28
Hispanic, race specified	1,251	1.04	1.15	945	1.81	1.29
Hispanic, race not specified	1,311	0.87	1.09	944	1.61	1.26
Asian	955	1.24	1.27	756	2.17	1.38
Hawaiian, Other Pacific Islander	170	0.99	1.11	132	1.91	1.39
American Indian, Alaska Native	232	0.52	0.91	175	1.41	1.33
More than one race, Non-Hispanic	379	1.16	1.33	259	2.36	1.36
Socioeconomic status						
1st quintile (lowest)	1,959	0.61	0.95	1,404	1.31	1.24
2nd quintile	2,228	0.84	1.11	1,765	1.80	1.33
3rd quintile	2,437	1.03	1.20	1,888	2.13	1.36
4th quintile	2,684	1.25	1.30	2,238	2.39	1.31
5th quintile (highest)	3,155	1.69	1.38	2,498	2.80	1.20
School type						
Public school	11,560	1.00	1.22	8,487	2.03	1.37
Private school	2,620	1.52	1.35	2,017	2.68	1.26

[1] Standard deviation.

NOTE: Table estimates are based on cross-sectional weights within each round (C5CW0, C6CW0). Estimates for third and fifth grade are counts of number right on the same set of items. The reading cluster was not tested in kindergarten/first grade. 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 and spring 2004.

www.manaraa.com

Table A14. Science: life science 5-item cluster score, third- and fifth-grade assessments (range of possible values: 0 to 5): School years 2001–02 and 2003–04

Characteristic	Round 5			Round 6		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	14,272	2.98	1.43	11,259	3.74	1.26
Sex						
Male	7,240	3.13	1.39	5,668	3.84	1.20
Female	7,032	2.81	1.44	5,591	3.64	1.32
Race/ethnicity						
White, Non-Hispanic	8,077	3.43	1.24	6,467	4.09	1.06
Black, Non-Hispanic	1,861	2.16	1.36	1,270	3.02	1.36
Hispanic, race specified	1,250	2.50	1.44	1,022	3.48	1.27
Hispanic, race not specified	1,315	2.22	1.43	1,079	3.24	1.42
Asian	950	2.93	1.50	783	3.82	1.26
Hawaiian, Other Pacific Islander	172	2.71	1.34	144	3.51	1.08
American Indian, Alaska Native	249	2.43	1.37	209	2.99	1.61
More than one race, Non-Hispanic	378	3.18	1.34	269	3.98	1.00
Socioeconomic status						
1st quintile (lowest)	1,982	2.07	1.37	1,705	2.92	1.43
2nd quintile	2,243	2.81	1.37	1,912	3.56	1.26
3rd quintile	2,440	3.10	1.30	1,991	3.90	1.03
4th quintile	2,684	3.39	1.26	2,308	4.08	1.03
5th quintile (highest)	3,153	3.80	1.15	2,534	4.36	0.93
School type						
Public school	11,600	2.93	1.43	9,185	3.69	1.29
Private school	2,623	3.37	1.31	2,052	4.10	1.04

[1] Standard deviation.

NOTE: Table estimates are based on cross-sectional weights within each round (C5CW0, C6CW0). Scores for third and fifth grade are counts of number right on the same set of items. Science was not tested in kindergarten/first grade. 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 and spring 2004.

www.manaraa.com

Table A15. Science: earth science 5-item cluster score, third- and fifth-grade assessments (range of possible values: 0 to 5): School years 2001–02 and 2003–04

Characteristic	Round 5			Round 6		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	14,298	2.69	1.37	11,263	3.37	1.22
Sex						
Male	7,245	2.82	1.38	5,671	3.48	1.21
Female	7,053	2.55	1.34	5,592	3.26	1.23
Race/ethnicity						
White, Non-Hispanic	8,095	3.07	1.26	6,467	3.71	1.07
Black, Non-Hispanic	1,859	2.00	1.31	1,271	2.65	1.24
Hispanic, race specified	1,256	2.26	1.37	1,022	3.06	1.20
Hispanic, race not specified	1,320	2.06	1.28	1,080	2.91	1.24
Asian	949	2.65	1.36	785	3.47	1.24
Hawaiian, Other Pacific Islander	172	2.41	1.37	144	3.18	1.28
American Indian, Alaska Native	249	2.32	1.38	209	2.76	1.47
More than one race, Non-Hispanic	378	2.85	1.26	269	3.68	1.13
Socioeconomic status						
1st quintile (lowest)	1,982	1.92	1.29	1,705	2.67	1.32
2nd quintile	2,246	2.50	1.32	1,915	3.24	1.16
3rd quintile	2,444	2.80	1.29	1,992	3.44	1.10
4th quintile	2,689	3.10	1.23	2,308	3.72	1.06
5th quintile (highest)	3,161	3.36	1.19	2,534	3.90	1.02
School type						
Public school	11,619	2.64	1.36	9,189	3.32	1.23
Private school	2,630	3.08	1.33	2,052	3.78	1.08

[1] Standard deviation.

NOTE: Table estimates are based on cross-sectional weights within each round (C5CW0, C6CW0). Scores for third and fifth grade are counts of number right on the same set of items. Science was not tested in kindergarten/first grade. 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 and spring 2004.

www.manaraa.com

Table A16. Science: physical science 5-item cluster score, third- and fifth-grade assessments (range of possible values: 0 to 5): School years 2001–02 and 2003–04

Characteristic	Round 5			Round 6		
	Number	Mean	SD[1]	Number	Mean	SD
Total sample	14,245	2.60	1.33	11,250	3.52	1.30
Sex						
Male	7,219	2.64	1.35	5,665	3.56	1.29
Female	7,026	2.56	1.31	5,585	3.48	1.32
Race/ethnicity						
White, Non-Hispanic	8,064	2.94	1.30	6,460	3.87	1.15
Black, Non-Hispanic	1,849	1.95	1.16	1,270	2.83	1.38
Hispanic, race specified	1,245	2.29	1.24	1,021	3.26	1.33
Hispanic, race not specified	1,319	2.06	1.24	1,080	2.98	1.30
Asian	950	2.78	1.35	783	3.66	1.24
Hawaiian, Other Pacific Islander	171	2.33	1.23	144	2.98	1.20
American Indian, Alaska Native	249	1.99	1.16	207	2.67	1.29
More than one race, Non-Hispanic	378	2.63	1.33	269	3.65	1.11
Socioeconomic status						
1st quintile (lowest)	1,974	1.88	1.17	1,701	2.72	1.33
2nd quintile	2,239	2.39	1.24	1,911	3.25	1.27
3rd quintile	2,440	2.66	1.27	1,991	3.69	1.17
4th quintile	2,679	2.94	1.27	2,307	3.87	1.21
5th quintile (highest)	3,149	3.36	1.23	2,533	4.21	0.96
School type						
Public school	11,574	2.56	1.33	9,177	3.47	1.32
Private school	2,622	2.94	1.33	2,051	3.96	1.12

[1] Standard deviation.

NOTE: Table estimates are based on cross-sectional weights within each round (C5CW0, C6CW0). Scores for third and fifth grade are counts of number right on the same set of items. Science was not tested in kindergarten/first grade. 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 and spring 2004.

www.manaraa.com

Table A17. Science: life science 7-item cluster score, third-grade assessment (range of possible values: 0 to 7): School year 2003–04

Characteristic	Round 6		
	Number	Mean	SD[1]
Total sample	11,246	4.71	1.73
Sex			
Male	5,660	4.84	1.68
Female	5,586	4.56	1.77
Race/ethnicity			
White, Non-Hispanic	6,459	5.21	1.51
Black, Non-Hispanic	1,268	3.64	1.71
Hispanic, race specified	1,020	4.31	1.67
Hispanic, race not specified	1,078	4.06	1.81
Asian	783	4.84	1.77
Hawaiian, Other Pacific Islander	144	4.29	1.54
American Indian, Alaska Native	209	3.56	2.06
More than one race, Non-Hispanic	269	4.97	1.37
Socioeconomic status			
1st quintile (lowest)	1,702	3.61	1.75
2nd quintile	1,910	4.42	1.70
3rd quintile	1,989	4.88	1.46
4th quintile	2,305	5.12	1.50
5th quintile (highest)	2,533	5.68	1.37
School type			
Public school	9,175	4.64	1.74
Private school	2,049	5.23	1.54

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 weight. There is no kindergarten/first-grade or third-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 2004.

www.manaraa.com

Table A18. Science: earth science 7-item cluster score, third-grade assessment (range of possible values: 0 to 7): School year 2003–04

Characteristic	Round 6		
	Number	Mean	SD[1]
Total sample	11,194	4.15	1.53
Sex			
Male	5,640	4.29	1.53
Female	5,554	4.00	1.52
Race/ethnicity			
White, Non-Hispanic	6,430	4.56	1.41
Black, Non-Hispanic	1,263	3.32	1.51
Hispanic, race specified	1,018	3.76	1.47
Hispanic, race not specified	1,077	3.52	1.45
Asian	775	4.34	1.58
Hawaiian, Other Pacific Islander	142	3.54	1.49
American Indian, Alaska Native	207	3.35	1.56
More than one race, Non-Hispanic	266	4.45	1.38
Socioeconomic status			
1st quintile (lowest)	1,686	3.29	1.54
2nd quintile	1,899	3.94	1.43
3rd quintile	1,983	4.18	1.38
4th quintile	2,298	4.57	1.37
5th quintile (highest)	2,523	4.93	1.37
School type			
Public school	9,126	4.07	1.53
Private school	2,047	4.72	1.44

[1] Standard deviation.

NOTE: Table estimates are based on C6CW0 weight. There is no kindergarten/first-grade or third-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 2004.

www.manaraa.com

Table A19. Science: physical science 7-item cluster score, third-grade assessment (range of possible values: 0 to 7): School year 2003–04

Characteristic	Round 6		
	Number	Mean	SD[1]
Total sample	11,243	4.32	1.74
Sex			
Male	5,663	4.46	1.74
Female	5,580	4.17	1.73
Race/ethnicity	,		
White, Non-Hispanic	6,459	4.80	1.60
Black, Non-Hispanic	1,269	3.30	1.65
Hispanic, race specified	1,018	3.97	1.71
Hispanic, race not specified	1,079	3.61	1.66
Asian	783	4.47	1.76
Hawaiian, Other Pacific Islander	144	3.64	1.68
American Indian, Alaska Native	206	3.29	1.60
More than one race, Non-Hispanic	269	4.62	1.50
Socioeconomic status			
1st quintile (lowest)	1,696	3.26	1.68
2nd quintile	1,911	3.93	1.66
3rd quintile	1,991	4.48	1.58
4th quintile	2,306	4.78	1.62
5th quintile (highest)	2,533	5.33	1.40
School type			
Public school	9,171	4.25	1.75
Private school	2,050	4.85	1.58

[1] Standard deviation.
NOTE: Table estimates are based on C6CW0 weight. There is no kindergarten/first-grade or third-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 2004.

www.manaraa.com

Table A20. Probability of proficiency, reading level 1: letter recognition (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.66	0.33	18,935	0.92	0.18	5,053	0.96	0.13	16,336	0.99	0.05	14,246	1.00	0.00	11,250	1.00	0.00
Sex																		
Male	8,984	0.63	0.34	9,688	0.91	0.19	2,556	0.95	0.14	8,349	0.99	0.06	7,204	1.00	0.00	5,660	1.00	0.00
Female	8,640	0.69	0.33	9,247	0.94	0.16	2,497	0.97	0.12	7,987	1.00	0.04	7,042	1.00	0.00	5,590	1.00	0.00
Race/ethnicity																		
White, Non-Hispanic	10,433	0.72	0.31	11,073	0.95	0.15	2,935	0.97	0.10	9,435	1.00	0.04	8,082	1.00	0.00	6,460	1.00	0.00
Black, Non-Hispanic	2,854	0.57	0.34	2,968	0.88	0.22	782	0.94	0.15	2,371	0.99	0.06	1,840	1.00	0.00	1,271	1.00	0.00
Hispanic, race specified	1,182	0.56	0.36	1,315	0.90	0.20	322	0.96	0.11	1,233	0.99	0.05	1,252	1.00	0.00	1,021	1.00	0.00
Hispanic, race not specified	1,195	0.50	0.36	1,423	0.86	0.24	377	0.91	0.20	1,335	0.99	0.06	1,314	1.00	0.00	1,077	1.00	0.00
Asian	896	0.76	0.30	1,088	0.96	0.12	257	0.98	0.08	1,042	0.99	0.05	956	1.00	0.00	785	1.00	0.00
Hawaiian, Other Pacific Islander	186	0.59	0.36	202	0.88	0.20	93	0.94	0.12	188	1.00	0.00	171	1.00	0.00	144	1.00	0.00
American Indian, Alaska Native	354	0.39	0.34	344	0.83	0.25	126	0.83	0.27	298	0.99	0.05	232	1.00	0.00	207	1.00	0.00
More than one race, Non-Hispanic	476	0.63	0.34	473	0.92	0.18	152	0.95	0.16	397	0.99	0.06	379	1.00	0.00	269	1.00	0.00
Socioeconomic status																		
1st quintile (lowest)	2,594	0.44	0.33	2,917	0.83	0.25	753	0.90	0.21	2,363	0.98	0.08	1,964	1.00	0.00	1,699	1.00	0.00
2nd quintile	3,271	0.58	0.33	3,503	0.90	0.20	925	0.95	0.15	2,796	0.99	0.05	2,230	1.00	0.00	1,912	1.00	0.00
3rd quintile	3,470	0.66	0.32	3,686	0.93	0.16	997	0.97	0.11	3,003	1.00	0.03	2,437	1.00	0.00	1,989	1.00	0.00
4th quintile	3,650	0.74	0.29	3,909	0.96	0.11	1,019	0.99	0.07	3,173	1.00	0.03	2,688	1.00	0.00	2,308	1.00	0.00
5th quintile (highest)	3,880	0.84	0.25	4,152	0.98	0.09	1,159	0.99	0.04	3,642	1.00	0.01	3,158	1.00	0.00	2,533	1.00	0.00
School type																		
Public school	13,736	0.63	0.34	14,578	0.91	0.19	3,809	0.96	0.14	12,998	0.99	0.05	11,575	1.00	0.00	9,177	1.00	0.00
Private school	3,888	0.81	0.26	4,357	0.96	0.12	1,042	0.99	0.05	3,279	1.00	0.03	2,623	1.00	0.00	2,051	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A21. Probability of proficiency, reading level 2: beginning sounds (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.29	0.32	18,935	0.68	0.33	5,053	0.81	0.27	16,336	0.96	0.13	14,246	1.00	0.00	11,250	1.00	0.00
Sex																		
Male	8,984	0.26	0.31	9,688	0.64	0.34	2,556	0.78	0.29	8,349	0.95	0.15	7,204	1.00	0.00	5,660	1.00	0.00
Female	8,640	0.32	0.33	9,247	0.72	0.31	2,497	0.84	0.25	7,987	0.97	0.11	7,042	1.00	0.00	5,590	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	0.23	9,435	0.98	0.11	8,082	1.00	0.00	6460	1.00	0.00
Black, Non-Hispanic	2,854	0.20	0.27	2,968	0.56	0.35	782	0.74	0.30	2,371	0.93	0.18	1,840	1.00	0.00	1,271	1.00	0.00
Hispanic, race specified	1,182	0.22	0.30	1,315	0.63	0.35	322	0.78	0.28	1,233	0.96	0.13	1,252	1.00	0.00	1,021	1.00	0.00
Hispanic, race not specified	1,195	0.17	0.26	1,423	0.55	0.36	377	0.68	0.34	1,335	0.94	0.15	1,314	1.00	0.00	1,077	1.00	0.00
Asian	896	0.40	0.36	1,088	0.78	0.28	257	0.85	0.23	1,042	0.97	0.12	956	1.00	0.00	785	1.00	0.00
Hawaiian, Other Pacific Islander	186	0.25	0.32	202	0.59	0.36	93	0.66	0.32	188	0.98	0.06	171	1.00	0.00	144	1.00	0.00
American Indian, Alaska Native	354	0.12	0.23	344	0.49	0.36	126	0.52	0.35	298	0.91	0.18	232	1.00	0.00	207	1.00	0.00
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	269	1.00	0.00
Socioeconomic status																		
1st quintile (lowest)	2,594	0.11	0.19	2,917	0.48	0.35	753	0.63	0.34	2,363	0.92	0.20	1,964	1.00	0.00	1,699	1.00	0.00
2nd quintile	3,271	0.20	0.26	3,503	0.61	0.34	925	0.76	0.29	2,796	0.96	0.14	2,230	1.00	0.00	1,912	1.00	0.00
3rd quintile	3,470	0.26	0.30	3,686	0.69	0.32	997	0.84	0.23	3,003	0.98	0.10	2,437	1.00	0.00	1,989	1.00	0.00
4th quintile	3,650	0.35	0.33	3,909	0.76	0.28	1,019	0.89	0.19	3,173	0.98	0.08	2,688	1.00	0.00	2,308	1.00	0.00
5th quintile (highest)	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	2,533	1.00	0.00
School type																		
Public school	13,736	0.26	0.31	14,578	0.65	0.34	3,809	0.80	0.28	12,998	0.96	0.14	11,575	1.00	0.00	9,177	1.00	0.00
Private school	3,888	0.44	0.35	4,357	0.80	0.27	1,042	0.92	0.16	3,279	0.99	0.07	2,623	1.00	0.00	2,051	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A22. Probability of proficiency, reading level 3: ending sounds (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.16	0.25	18,935	0.49	0.34	5,053	0.65	0.32	16,336	0.92	0.19	14,246	1.00	0.01	11,250	1.00	0.00
Sex																		
Male	8,984	0.14	0.24	9,688	0.45	0.35	2,556	0.61	0.33	8,349	0.90	0.21	7,204	1.00	0.01	5,660	1.00	0.00
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	5,590	1.00	0.00
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	0.15	8,082	1.00	0.01	6,460	1.00	0.00
Black, Non-Hispanic	2,854	0.10	0.20	2,968	0.37	0.34	782	0.55	0.34	2,371	0.86	0.25	1,840	1.00	0.01	1,271	1.00	0.00
Hispanic, race specified	1,182	0.12	0.22	1,315	0.44	0.34	322	0.61	0.34	1,233	0.90	0.20	1,252	1.00	0.01	1,021	1.00	0.00
Hispanic, race not specified	1,195	0.08	0.17	1,423	0.36	0.33	377	0.49	0.34	1,335	0.87	0.22	1,314	1.00	0.01	1,077	1.00	0.00
Asian	896	0.25	0.32	1,088	0.60	0.34	257	0.71	0.32	1,042	0.94	0.17	956	1.00	0.00	785	1.00	0.00
Hawaiian, Other Pacific Islander	186	0.14	0.24	202	0.41	0.35	93	0.46	0.34	188	0.92	0.14	171	1.00	0.02	144	1.00	0.00
American Indian, Alaska Native	354	0.06	0.15	344	0.31	0.31	126	0.32	0.31	298	0.80	0.27	232	0.99	0.02	207	1.00	0.00
More than one race, Non-Hispanic	476	0.16	0.26	473	0.46	0.33	152	0.65	0.31	397	0.92	0.19	379	1.00	0.01	269	1.00	0.00
Socioeconomic status																		
1st quintile (lowest)	2,594	0.05	0.12	2,917	0.28	0.29	753	0.43	0.33	2,363	0.83	0.26	1,964	0.99	0.02	1,699	1.00	0.00
2nd quintile	3,271	0.09	0.18	3,503	0.41	0.33	925	0.57	0.33	2,796	0.90	0.20	2,230	1.00	0.01	1,912	1.00	0.00
3rd quintile	3,470	0.13	0.22	3,686	0.49	0.33	997	0.68	0.30	3,003	0.94	0.15	2,437	1.00	0.01	1,989	1.00	0.00
4th quintile	3,650	0.20	0.26	3,909	0.57	0.32	1,019	0.74	0.27	3,173	0.96	0.12	2,688	1.00	0.01	2,308	1.00	0.00
5th quintile (highest)	3,880	0.31	0.32	4,152	0.68	0.30	1,159	0.81	0.24	3,642	0.97	0.08	3,158	1.00	0.00	2,533	1.00	0.00
School type																		
Public school	13,736	0.14	0.23	14,578	0.46	0.34	3,809	0.63	0.33	12,998	0.91	0.20	11,575	1.00	0.01	9,177	1.00	0.00
Private school	3,888	0.27	0.31	4,357	0.63	0.32	1,042	0.80	0.23	3,279	0.97	0.11	2,623	1.00	0.00	2,051	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A23. Probability of proficiency, reading level 4: sight words (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.03	0.13	18,935	0.15	0.26	5,053	0.26	0.33	16,336	0.74	0.33	14,246	0.99	0.07	11,250	1.00	0.01
Sex																		
Male	8,984	0.03	0.13	9,688	0.13	0.25	2,556	0.24	0.31	8,349	0.70	0.35	7,204	0.98	0.08	5,660	1.00	0.01
Female	8,640	0.03	0.12	9,247	0.16	0.27	2,497	0.29	0.34	7,987	0.77	0.31	7,042	0.99	0.06	5,590	1.00	0.01
Race/ethnicity																		
White, Non-Hispanic	10,433	0.04	0.14	11,073	0.17	0.27	2,935	0.31	0.34	9,435	0.80	0.29	8,082	0.99	0.05	6,460	1.00	0.01
Black, Non-Hispanic	2,854	0.01	0.09	2,968	0.10	0.21	782	0.19	0.29	2,371	0.63	0.38	1,840	0.97	0.09	1,271	1.00	0.02
Hispanic, race specified	1,182	0.02	0.10	1,315	0.12	0.23	322	0.23	0.30	1,233	0.68	0.36	1,252	0.98	0.08	1,021	1.00	0.01
Hispanic, race not specified	1,195	0.01	0.07	1,423	0.08	0.19	377	0.15	0.26	1,335	0.61	0.37	1,314	0.97	0.10	1,077	1.00	0.02
Asian	896	0.08	0.24	1,088	0.26	0.35	257	0.41	0.42	1,042	0.81	0.31	956	0.99	0.04	785	1.00	0.01
Hawaiian, Other Pacific Islander	186	0.03	0.13	202	0.13	0.23	93	0.15	0.28	188	0.68	0.35	171	0.98	0.09	144	1.00	0.00
American Indian, Alaska Native	354	0.01	0.05	344	0.06	0.15	126	0.07	0.17	298	0.48	0.38	232	0.95	0.14	207	0.99	0.02
More than one race, Non-Hispanic	476	0.04	0.16	473	0.13	0.26	152	0.27	0.34	397	0.76	0.31	379	0.99	0.05	269	1.00	0.00
Socioeconomic status																		
1st quintile (lowest)	2,594	0.00	0.05	2,917	0.05	0.14	753	0.10	0.21	2,363	0.54	0.38	1,964	0.96	0.12	1,699	0.99	0.02
2nd quintile	3,271	0.01	0.08	3,503	0.10	0.21	925	0.18	0.28	2,796	0.70	0.35	2,230	0.98	0.08	1,912	1.00	0.01
3rd quintile	3,470	0.02	0.09	3,686	0.13	0.23	997	0.26	0.32	3,003	0.77	0.31	2,437	0.99	0.04	1,989	1.00	0.01
4th quintile	3,650	0.03	0.13	3,909	0.18	0.27	1,019	0.32	0.33	3,173	0.82	0.26	2,688	1.00	0.03	2,308	1.00	0.01
5th quintile (highest)	3,880	0.08	0.20	4,152	0.28	0.33	1,159	0.44	0.37	3,642	0.88	0.23	3,158	1.00	0.01	2,533	1.00	0.00
School type																		
Public school	13,736	0.02	0.11	14,578	0.13	0.24	3,809	0.24	0.32	12,998	0.72	0.34	11,575	0.98	0.07	9,177	1.00	0.01
Private school	3,888	0.06	0.18	4,357	0.24	0.32	1,042	0.40	0.36	3,279	0.86	0.25	2,623	1.00	0.02	2,051	1.00	0.01

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A24. Probability of proficiency, reading level 5: words in context (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.01	0.08	18,935	0.06	0.15	5,053	0.12	0.22	16,336	0.45	0.33	14,246	0.90	0.17	11,250	0.97	0.08
Sex																		
Male	8,984	0.01	0.08	9,688	0.05	0.15	2,556	0.11	0.22	8,349	0.42	0.33	7,204	0.89	0.19	5,660	0.96	0.09
Female	8,640	0.01	0.07	9,247	0.07	0.16	2,497	0.14	0.23	7,987	0.49	0.32	7,042	0.92	0.15	5,590	0.97	0.07
Race/ethnicity																		
White, Non-Hispanic	10,433	0.02	0.08	11,073	0.07	0.16	2,935	0.14	0.24	9,435	0.51	0.32	8,082	0.94	0.13	6,460	0.98	0.06
Black, Non-Hispanic	2,854	0.01	0.05	2,968	0.04	0.10	782	0.08	0.17	2,371	0.35	0.31	1,840	0.85	0.20	1,271	0.95	0.10
Hispanic, race specified	1,182	0.01	0.07	1,315	0.05	0.13	322	0.09	0.18	1,233	0.40	0.32	1,252	0.88	0.18	1,021	0.96	0.08
Hispanic, race not specified	1,195	0.00	0.04	1,423	0.03	0.09	377	0.06	0.14	1,335	0.32	0.30	1,314	0.83	0.22	1,077	0.95	0.10
Asian	896	0.05	0.16	1,088	0.13	0.24	257	0.25	0.33	1,042	0.56	0.34	956	0.93	0.12	785	0.97	0.08
Hawaiian, Other Pacific Islander	186	0.01	0.07	202	0.04	0.09	93	0.07	0.17	188	0.39	0.32	171	0.89	0.15	144	0.97	0.05
American Indian, Alaska Native	354	0.00	0.02	344	0.02	0.07	126	0.03	0.09	298	0.24	0.27	232	0.78	0.25	207	0.92	0.13
More than one race, Non-Hispanic	476	0.02	0.11	473	0.06	0.17	152	0.12	0.21	397	0.47	0.32	379	0.91	0.16	269	0.98	0.04
Socioeconomic status																		
1st quintile (lowest)	2,594	0.00	0.03	2,917	0.02	0.06	753	0.04	0.10	2,363	0.27	0.27	1,964	0.80	0.24	1,699	0.92	0.13
2nd quintile	3,271	0.01	0.05	3,503	0.04	0.11	925	0.07	0.16	2,796	0.40	0.31	2,230	0.89	0.18	1,912	0.96	0.07
3rd quintile	3,470	0.01	0.05	3,686	0.05	0.12	997	0.12	0.22	3,003	0.46	0.31	2,437	0.92	0.13	1,989	0.98	0.05
4th quintile	3,650	0.01	0.08	3,909	0.07	0.16	1,019	0.15	0.24	3,173	0.53	0.31	2,688	0.95	0.11	2,308	0.98	0.05
5th quintile (highest)	3,880	0.04	0.13	4,152	0.13	0.22	1,159	0.22	0.30	3,642	0.63	0.31	3,158	0.97	0.07	2,533	0.99	0.02
School type																		
Public school	13,736	0.01	0.07	14,578	0.05	0.14	3,809	0.11	0.21	12,998	0.43	0.32	11,575	0.90	0.17	9,177	0.97	0.08
Private school	3,888	0.03	0.11	4,357	0.11	0.20	1,042	0.20	0.27	3,279	0.60	0.31	2,623	0.96	0.08	2,051	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A25. Probability of proficiency, reading level 6: literal inference (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.00	0.03	18,935	0.01	0.07	5,053	0.03	0.12	16,336	0.16	0.23	14,246	0.68	0.30	11,250	0.86	0.21
Sex																		
Male	8,984	0.00	0.04	9,688	0.01	0.07	2,556	0.03	0.12	8,349	0.15	0.22	7,204	0.65	0.31	5,660	0.84	0.22
Female	8,640	0.00	0.03	9,247	0.01	0.07	2,497	0.04	0.12	7,987	0.17	0.23	7,042	0.71	0.28	5,590	0.87	0.19
Race/ethnicity																		
White, Non-Hispanic	10,433	0.00	0.04	11,073	0.02	0.08	2,935	0.04	0.14	9,435	0.19	0.25	8,082	0.75	0.26	6,460	0.90	0.17
Black, Non-Hispanic	2,854	0.00	0.02	2,968	0.00	0.03	782	0.01	0.07	2,371	0.09	0.16	1,840	0.55	0.30	1,271	0.78	0.24
Hispanic, race specified	1,182	0.00	0.04	1,315	0.01	0.05	322	0.02	0.09	1,233	0.12	0.20	1,252	0.61	0.31	1,021	0.83	0.22
Hispanic, race not specified	1,195	0.00	0.01	1,423	0.00	0.03	377	0.01	0.06	1,335	0.09	0.16	1,314	0.53	0.32	1,077	0.78	0.24
Asian	896	0.01	0.06	1,088	0.03	0.11	257	0.09	0.19	1,042	0.23	0.27	956	0.72	0.27	785	0.89	0.19
Hawaiian, Other Pacific Islander	186	0.00	0.02	202	0.00	0.03	93	0.01	0.05	188	0.12	0.20	171	0.61	0.29	144	0.84	0.19
American Indian, Alaska Native	354	0.00	0.00	344	0.00	0.01	126	0.00	0.02	298	0.05	0.13	232	0.45	0.32	207	0.70	0.29
More than one race, Non-Hispanic	476	0.01	0.04	473	0.02	0.09	152	0.03	0.10	397	0.16	0.24	379	0.69	0.29	269	0.90	0.17
Socioeconomic status																		
1st quintile (lowest)	2,594	0.00	0.01	2,917	0.00	0.02	753	0.01	0.04	2,363	0.06	0.12	1,964	0.47	0.31	1,699	0.71	0.27
2nd quintile	3,271	0.00	0.02	3,503	0.01	0.05	925	0.01	0.07	2,796	0.12	0.19	2,230	0.63	0.29	1,912	0.83	0.21
3rd quintile	3,470	0.00	0.01	3,686	0.01	0.04	997	0.03	0.11	3,003	0.15	0.21	2,437	0.71	0.27	1,989	0.89	0.16
4th quintile	3,650	0.00	0.04	3,909	0.01	0.07	1,019	0.04	0.13	3,173	0.19	0.24	2,688	0.77	0.24	2,308	0.91	0.15
5th quintile (highest)	3,880	0.01	0.06	4,152	0.03	0.12	1,159	0.08	0.18	3,642	0.29	0.29	3,158	0.86	0.19	2,533	0.96	0.09
School type																		
Public school	13,736	0.00	0.03	14,578	0.01	0.06	3,809	0.03	0.11	12,998	0.14	0.22	11,575	0.66	0.30	9,177	0.85	0.21
Private school	3,888	0.01	0.05	4,357	0.02	0.10	1,042	0.05	0.15	3,279	0.25	0.27	2,623	0.80	0.23	2,051	0.92	0.15

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A26. Probability of proficiency, reading level 7: extrapolation (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	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	0.05	16,336	0.03	0.11	14,246	0.42	0.38	11,250	0.69	0.35
Sex																		
Male	8,984	0.00	0.01	9,688	0.00	0.03	2,556	0.01	0.05	8,349	0.03	0.11	7,204	0.39	0.37	5,660	0.67	0.36
Female	8,640	0.00	0.01	9,247	0.00	0.02	2,497	0.01	0.05	7,987	0.03	0.11	7,042	0.44	0.38	5,590	0.72	0.34
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	0.04	0.13	8,082	0.51	0.38	6,460	0.78	0.31
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	0.25	0.31	1,271	0.54	0.36
Hispanic, race specified	1,182	0.00	0.01	1,315	0.00	0.02	322	0.00	0.06	1,233	0.02	0.08	1,252	0.33	0.36	1,021	0.63	0.37
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	0.06	1,314	0.24	0.31	1,077	0.53	0.37
Asian	896	0.00	0.01	1,088	0.00	0.03	257	0.02	0.10	1,042	0.06	0.15	956	0.46	0.38	785	0.76	0.32
Hawaiian, Other Pacific Islander	186	0.00	0.00	202	0.00	0.00	93	0.00	0.00	188	0.02	0.09	171	0.31	0.34	144	0.63	0.35
American Indian, Alaska Native	354	0.00	0.00	344	0.00	0.00	126	0.00	0.00	298	0.01	0.06	232	0.18	0.29	207	0.46	0.41
More than one race, Non-Hispanic	476	0.00	0.01	473	0.00	0.03	152	0.00	0.02	397	0.04	0.14	379	0.42	0.36	269	0.78	0.31
Socioeconomic status																		
1st quintile (lowest)	2,594	0.00	0.00	2,917	0.00	0.01	753	0.00	0.02	2,363	0.00	0.03	1,964	0.18	0.27	1,699	0.43	0.36
2nd 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	0.33	0.34	1,912	0.63	0.35
3rd quintile	3,470	0.00	0.00	3,686	0.00	0.01	997	0.00	0.03	3,003	0.02	0.09	2,437	0.42	0.36	1,989	0.74	0.31
4th 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	0.52	0.37	2,308	0.81	0.29
5th quintile (highest)	3,880	0.00	0.01	4,152	0.01	0.05	1,159	0.02	0.09	3,642	0.08	0.18	3,158	0.68	0.34	2,533	0.90	0.20
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	0.10	11,575	0.40	0.37	9,177	0.68	0.36
Private school	3,888	0.00	0.01	4,357	0.00	0.04	1,042	0.01	0.07	3,279	0.06	0.15	2,623	0.57	0.37	2,051	0.83	0.28

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A27. Probability of proficiency, reading level 8: evaluation (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.00	0.01	18,935	0.00	0.01	5,053	0.01	0.03	16,336	0.03	0.06	14,246	0.24	0.20	11,250	0.44	0.27
Sex																		
Male	8,984	0.00	0.01	9,688	0.00	0.01	2,556	0.01	0.03	8,349	0.03	0.06	7,204	0.22	0.20	5,660	0.42	0.27
Female	8,640	0.00	0.01	9,247	0.00	0.01	2,497	0.01	0.03	7,987	0.03	0.06	7,042	0.25	0.21	5,590	0.45	0.27
Race/ethnicity																		
White, Non-Hispanic	10,433	0.00	0.01	11,073	0.00	0.02	2,935	0.01	0.03	9,435	0.04	0.07	8,082	0.29	0.21	6,460	0.50	0.27
Black, Non-Hispanic	2,854	0.00	0.00	2,968	0.00	0.01	782	0.00	0.01	2,371	0.02	0.04	1,840	0.15	0.14	1,271	0.31	0.23
Hispanic, race specified	1,182	0.00	0.01	1,315	0.00	0.01	322	0.00	0.03	1,233	0.02	0.04	1,252	0.19	0.18	1,021	0.38	0.26
Hispanic, race not specified	1,195	0.00	0.00	1,423	0.00	0.01	377	0.00	0.01	1,335	0.02	0.03	1,314	0.14	0.15	1,077	0.30	0.23
Asian	896	0.00	0.01	1,088	0.01	0.02	257	0.02	0.05	1,042	0.05	0.08	956	0.26	0.21	785	0.48	0.26
Hawaiian, Other Pacific Islander	186	0.00	0.00	202	0.00	0.00	93	0.00	0.01	188	0.02	0.05	171	0.18	0.16	144	0.39	0.26
American Indian, Alaska Native	354	0.00	0.00	344	0.00	0.00	126	0.00	0.00	298	0.01	0.03	232	0.12	0.14	207	0.29	0.26
More than one race, Non-Hispanic	476	0.00	0.01	473	0.00	0.02	152	0.01	0.02	397	0.04	0.07	379	0.24	0.21	269	0.49	0.25
Socioeconomic status																		
1st quintile (lowest)	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	0.12	0.13	1,699	0.25	0.21
2nd quintile	3,271	0.00	0.00	3,503	0.00	0.01	925	0.00	0.02	2,796	0.02	0.04	2,230	0.19	0.17	1,912	0.37	0.24
3rd quintile	3,470	0.00	0.00	3,686	0.00	0.01	997	0.01	0.02	3,003	0.03	0.05	2,437	0.23	0.18	1,989	0.45	0.25
4th quintile	3,650	0.00	0.01	3,909	0.00	0.01	1,019	0.01	0.03	3,173	0.04	0.06	2,688	0.29	0.21	2,308	0.51	0.25
5th quintile (highest)	3,880	0.00	0.01	4,152	0.01	0.03	1,159	0.02	0.05	3,642	0.06	0.09	3,158	0.38	0.22	2,533	0.63	0.24
School type																		
Public school	13,736	0.00	0.01	14,578	0.00	0.01	3,809	0.01	0.03	12,998	0.03	0.05	11,575	0.23	0.20	9,177	0.42	0.27
Private school	3,888	0.00	0.01	4,357	0.01	0.02	1,042	0.01	0.04	3,279	0.05	0.08	2,623	0.32	0.22	2,051	0.55	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A28. Probability of proficiency, reading level 9: evaluating nonfiction (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	17,624	0.00	0.00	18,935	0.00	0.00	5,053	0.00	0.00	16,336	0.00	0.00	14,246	0.01	0.04	11,250	0.07	0.18
Sex																		
Male	8,984	0.00	0.00	9,688	0.00	0.00	2,556	0.00	0.00	8,349	0.00	0.00	7,204	0.01	0.05	5,660	0.07	0.18
Female	8,640	0.00	0.00	9,247	0.00	0.00	2,497	0.00	0.00	7,987	0.00	0.00	7,042	0.01	0.04	5,590	0.07	0.18
Race/ethnicity																		
White, Non-Hispanic	10,433	0.00	0.00	11,073	0.00	0.00	2,935	0.00	0.00	9,435	0.00	0.00	8,082	0.01	0.05	6,460	0.10	0.21
Black, Non-Hispanic	2,854	0.00	0.00	2,968	0.00	0.00	782	0.00	0.00	2,371	0.00	0.00	1,840	0.00	0.01	1,271	0.02	0.09
Hispanic, race specified	1,182	0.00	0.00	1,315	0.00	0.00	322	0.00	0.00	1,233	0.00	0.00	1,252	0.00	0.03	1,021	0.04	0.13
Hispanic, race not specified	1,195	0.00	0.00	1,423	0.00	0.00	377	0.00	0.00	1,335	0.00	0.00	1,314	0.00	0.01	1,077	0.02	0.09
Asian	896	0.00	0.00	1,088	0.00	0.00	257	0.00	0.00	1,042	0.00	0.00	956	0.01	0.03	785	0.07	0.17
Hawaiian, Other Pacific Islander	186	0.00	0.00	202	0.00	0.00	93	0.00	0.00	188	0.00	0.00	171	0.00	0.01	144	0.05	0.14
American Indian, Alaska Native	354	0.00	0.00	344	0.00	0.00	126	0.00	0.00	298	0.00	0.00	232	0.00	0.01	207	0.03	0.10
More than one race, Non-Hispanic	476	0.00	0.00	473	0.00	0.00	152	0.00	0.00	397	0.00	0.00	379	0.01	0.05	269	0.08	0.19
Socioeconomic status																		
1st quintile (lowest)	2,594	0.00	0.00	2,917	0.00	0.00	753	0.00	0.00	2,363	0.00	0.00	1,964	0.00	0.01	1,699	0.01	0.07
2nd quintile	3,271	0.00	0.00	3,503	0.00	0.00	925	0.00	0.00	2,796	0.00	0.00	2,230	0.00	0.03	1,912	0.03	0.11
3rd quintile	3,470	0.00	0.00	3,686	0.00	0.00	997	0.00	0.00	3,003	0.00	0.00	2,437	0.00	0.03	1,989	0.05	0.14
4th quintile	3,650	0.00	0.00	3,909	0.00	0.00	1,019	0.00	0.00	3,173	0.00	0.00	2,688	0.01	0.05	2,308	0.09	0.19
5th quintile (highest)	3,880	0.00	0.00	4,152	0.00	0.00	1,159	0.00	0.00	3,642	0.00	0.00	3,158	0.02	0.08	2,533	0.17	0.27
School type																		
Public school	13,736	0.00	0.00	14,578	0.00	0.00	3,809	0.00	0.00	12,998	0.00	0.00	11,575	0.01	0.04	9,177	0.06	0.17
Private school	3,888	0.00	0.00	4,357	0.00	0.00	1,042	0.00	0.00	3,279	0.00	0.00	2,623	0.02	0.07	2,051	0.13	0.24

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A29. Probability of proficiency, mathematics level 1: count, number, shape (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round\ 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.91	0.18	19,647	0.99	0.06	5,226	0.99	0.04	16,641	1.00	0.02	14,349	1.00	0.00	11,266	1.00	0.00
Sex																		
Male	9,479	0.91	0.19	10,041	0.99	0.07	2,644	0.99	0.04	8,506	1.00	0.02	7,277	1.00	0.00	5,672	1.00	0.00
Female	9,156	0.92	0.17	9,606	0.99	0.06	2,582	1.00	0.04	8,135	1.00	0.01	7,072	1.00	0.00	5,594	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	6,467	1.00	0.00
Black, Non-Hispanic	2,855	0.88	0.21	2,962	0.98	0.08	781	0.99	0.05	2,371	1.00	0.03	1,871	1.00	0.00	1,275	1.00	0.00
Hispanic, race specified	1,588	0.86	0.23	1,624	0.98	0.07	389	0.99	0.04	1,354	1.00	0.00	1,260	1.00	0.00	1,022	1.00	0.00
Hispanic, race not specified	1,800	0.81	0.26	1,834	0.97	0.10	486	0.99	0.05	1,518	1.00	0.01	1,324	1.00	0.00	1,080	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	785	1.00	0.00
Hawaiian, Other Pacific Islander	187	0.89	0.21	202	0.99	0.05	93	1.00	0.01	188	1.00	0.00	172	1.00	0.00	144	1.00	0.00
American Indian, Alaska Native	354	0.80	0.26	345	0.98	0.08	126	0.97	0.11	298	1.00	0.00	250	1.00	0.00	208	1.00	0.00
More than one race, Non-Hispanic	473	0.93	0.14	472	0.99	0.05	151	1.00	0.02	397	1.00	0.02	380	1.00	0.00	269	1.00	0.00
Socioeconomic status																		
1st quintile (lowest)	3,269	0.81	0.26	3,426	0.97	0.10	895	0.99	0.06	2,572	1.00	0.02	2,001	1.00	0.00	1,708	1.00	0.00
2nd quintile	3,429	0.89	0.20	3,607	0.98	0.07	942	0.99	0.05	2,839	1.00	0.03	2,250	1.00	0.00	1,915	1.00	0.00
3rd quintile	3,546	0.94	0.15	3,721	0.99	0.05	1,001	1.00	0.03	3,017	1.00	0.00	2,452	1.00	0.00	1,991	1.00	0.00
4th quintile	3,676	0.96	0.12	3,921	1.00	0.04	1,023	1.00	0.02	3,178	1.00	0.01	2,693	1.00	0.00	2,308	1.00	0.00
5th quintile (highest)	3,893	0.98	0.08	4,161	1.00	0.03	1,158	1.00	0.00	3,644	1.00	0.01	3,163	1.00	0.00	2,534	1.00	0.00
School type																		
Public school	14,701	0.90	0.19	15,259	0.99	0.07	3,971	0.99	0.04	13,292	1.00	0.02	11,670	1.00	0.00	9,193	1.00	0.00
Private school	3,934	0.97	0.10	4,388	1.00	0.05	1,043	1.00	0.00	3,286	1.00	0.02	2,631	1.00	0.00	2,052	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A30. Probability of proficiency, mathematics level 2: relative size (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.53	0.35	19,647	0.83	0.24	5,226	0.91	0.18	16,641	0.98	0.08	14,349	1.00	0.00	11,266	1.00	0.00
Sex																		
Male	9,479	0.53	0.35	10,041	0.83	0.25	2,644	0.90	0.20	8,506	0.98	0.09	7,277	1.00	0.00	5,672	1.00	0.00
Female	9,156	0.53	0.34	9,606	0.84	0.24	2,582	0.92	0.16	8,135	0.99	0.07	7,072	1.00	0.00	5,594	1.00	0.00
Race/ethnicity																		
White, Non-Hispanic	10,433	0.63	0.33	11,071	0.89	0.19	2,935	0.95	0.13	9,436	0.99	0.07	8,116	1.00	0.00	6,467	1.00	0.00
Black, Non-Hispanic	2,855	0.41	0.32	2,962	0.74	0.28	781	0.86	0.22	2,371	0.97	0.12	1,871	1.00	0.00	1,275	1.00	0.00
Hispanic, race specified	1,588	0.40	0.34	1,624	0.77	0.28	389	0.89	0.20	1,354	0.98	0.08	1,260	1.00	0.00	1,022	1.00	0.00
Hispanic, race not specified	1,800	0.32	0.32	1,834	0.71	0.30	486	0.83	0.25	1,518	0.98	0.08	1,324	1.00	0.00	1,080	1.00	0.00
Asian	897	0.66	0.32	1,088	0.90	0.17	256	0.94	0.12	1,042	0.99	0.05	956	1.00	0.00	785	1.00	0.00
Hawaiian, Other Pacific Islander	187	0.46	0.33	202	0.78	0.27	93	0.89	0.17	188	0.99	0.04	172	1.00	0.00	144	1.00	0.00
American Indian, Alaska Native	354	0.34	0.33	345	0.73	0.29	126	0.77	0.30	298	0.97	0.09	250	1.00	0.00	208	1.00	0.00
More than one race, Non-Hispanic	473	0.52	0.34	472	0.85	0.22	151	0.91	0.19	397	0.98	0.09	380	1.00	0.00	269	1.00	0.00
Socioeconomic status																		
1st quintile (lowest)	3,269	0.30	0.30	3,426	0.69	0.30	895	0.81	0.26	2,572	0.97	0.11	2,001	1.00	0.00	1,708	1.00	0.00
2nd quintile	3,429	0.45	0.33	3,607	0.80	0.26	942	0.90	0.20	2,839	0.98	0.10	2,250	1.00	0.00	1,915	1.00	0.00
3rd quintile	3,546	0.55	0.33	3,721	0.86	0.21	1,001	0.94	0.15	3,017	0.99	0.07	2,452	1.00	0.00	1,991	1.00	0.00
4th quintile	3,676	0.64	0.31	3,921	0.90	0.17	1,023	0.96	0.11	3,178	0.99	0.05	2,693	1.00	0.00	2,308	1.00	0.00
5th quintile (highest)	3,893	0.75	0.28	4,161	0.94	0.14	1,158	0.98	0.07	3,644	1.00	0.03	3,163	1.00	0.00	2,534	1.00	0.00
School type																		
Public school	14,701	0.50	0.35	15,259	0.82	0.25	3,971	0.91	0.19	13,292	0.98	0.08	11,670	1.00	0.00	9,193	1.00	0.00
Private school	3,934	0.70	0.30	4,388	0.92	0.16	1,043	0.98	0.06	3,286	1.00	0.04	2,631	1.00	0.00	2,052	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A31. Probability of proficiency, mathematics level 3: ordinality, sequence (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.19	0.30	19,647	0.52	0.39	5,226	0.72	0.35	16,641	0.94	0.19	14,349	1.00	0.01	11,266	1.00	0.00
Sex																		
Male	9,479	0.19	0.30	10,041	0.52	0.39	2,644	0.71	0.36	8,506	0.93	0.19	7,277	1.00	0.01	5,672	1.00	0.00
Female	9,156	0.18	0.29	9,606	0.53	0.39	2,582	0.73	0.34	8,135	0.94	0.18	7,072	1.00	0.01	5,594	1.00	0.00
Race/ethnicity																		
White, Non-Hispanic	10,433	0.25	0.33	11,071	0.63	0.37	2,935	0.81	0.30	9,436	0.96	0.15	8,116	1.00	0.00	6,467	1.00	0.00
Black, Non-Hispanic	2,855	0.09	0.19	2,962	0.36	0.37	781	0.59	0.39	2,371	0.88	0.26	1,871	1.00	0.01	1,275	1.00	0.00
Hispanic, race specified	1,588	0.11	0.23	1,624	0.42	0.38	389	0.67	0.37	1,354	0.91	0.22	1,260	1.00	0.01	1,022	1.00	0.00
Hispanic, race not specified	1,800	0.07	0.18	1,834	0.33	0.36	486	0.53	0.40	1,518	0.91	0.21	1,324	1.00	0.02	1,080	1.00	0.00
Asian	897	0.30	0.36	1,088	0.63	0.38	256	0.77	0.34	1,042	0.95	0.15	956	1.00	0.00	785	1.00	0.00
Hawaiian, Other Pacific Islander	187	0.11	0.23	202	0.41	0.37	93	0.57	0.36	188	0.91	0.19	172	1.00	0.01	144	1.00	0.00
American Indian, Alaska Native	354	0.08	0.19	345	0.35	0.36	126	0.44	0.40	298	0.90	0.22	250	1.00	0.01	208	1.00	0.00
More than one race, Non-Hispanic	473	0.17	0.29	472	0.51	0.38	151	0.73	0.34	397	0.94	0.18	380	1.00	0.00	269	1.00	0.00
Socioeconomic status																		
1st quintile (lowest)	3,269	0.05	0.15	3,426	0.29	0.33	895	0.47	0.39	2,572	0.87	0.25	2,001	1.00	0.01	1,708	1.00	0.00
2nd quintile	3,429	0.12	0.23	3,607	0.46	0.38	942	0.67	0.36	2,839	0.92	0.21	2,250	1.00	0.01	1,915	1.00	0.00
3rd quintile	3,546	0.17	0.27	3,721	0.54	0.37	1,001	0.77	0.31	3,017	0.95	0.16	2,452	1.00	0.01	1,991	1.00	0.00
4th quintile	3,676	0.24	0.32	3,921	0.63	0.36	1,023	0.83	0.27	3,178	0.97	0.12	2,693	1.00	0.00	2,308	1.00	0.00
5th quintile (highest)	3,893	0.37	0.37	4,161	0.75	0.32	1,158	0.89	0.23	3,644	0.99	0.08	3,163	1.00	0.00	2,534	1.00	0.00
School type																		
Public school	14,701	0.16	0.28	15,259	0.50	0.39	3,971	0.70	0.36	13,292	0.93	0.20	11,670	1.00	0.01	9,193	1.00	0.00
Private school	3,934	0.31	0.35	4,388	0.68	0.35	1,043	0.88	0.22	3,286	0.98	0.09	2,631	1.00	0.00	2,052	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A32. Probability of proficiency, mathematics level 4: add/subtract (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.03	0.11	19,647	0.15	0.25	5,226	0.31	0.33	16,641	0.70	0.33	14,349	0.97	0.10	11,266	0.99	0.03
Sex																		
Male	9,479	0.04	0.13	10,041	0.16	0.26	2,644	0.32	0.34	8,506	0.70	0.33	7,277	0.97	0.10	5,672	0.99	0.03
Female	9,156	0.03	0.10	9,606	0.15	0.23	2,582	0.30	0.31	8,135	0.70	0.32	7,072	0.97	0.09	5,594	0.99	0.03
Race/ethnicity																		
White, Non-Hispanic	10,433	0.05	0.13	11,071	0.20	0.27	2,935	0.38	0.34	9,436	0.77	0.29	8,116	0.98	0.07	6,467	1.00	0.02
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	0.94	0.13	1,275	0.99	0.04
Hispanic, race specified	1,588	0.01	0.07	1,624	0.10	0.19	389	0.25	0.29	1,354	0.64	0.34	1,260	0.96	0.11	1,022	0.99	0.03
Hispanic, race not specified	1,800	0.01	0.04	1,834	0.07	0.16	486	0.16	0.24	1,518	0.58	0.33	1,324	0.95	0.12	1,080	0.99	0.03
Asian	897	0.08	0.19	1,088	0.23	0.30	256	0.41	0.36	1,042	0.73	0.32	956	0.98	0.07	785	0.99	0.03
Hawaiian, Other Pacific Islander	187	0.02	0.09	202	0.09	0.18	93	0.16	0.24	188	0.56	0.33	172	0.96	0.12	144	1.00	0.01
American Indian, Alaska Native	354	0.01	0.05	345	0.08	0.16	126	0.12	0.21	298	0.53	0.35	250	0.94	0.11	208	0.99	0.04
More than one race, Non-Hispanic	473	0.03	0.11	472	0.14	0.23	151	0.27	0.29	397	0.70	0.33	380	0.97	0.10	269	0.99	0.03
Socioeconomic status																		
1st quintile (lowest)	3,269	0.00	0.04	3,426	0.05	0.13	895	0.14	0.23	2,572	0.52	0.35	2,001	0.92	0.15	1,708	0.98	0.05
2nd quintile	3,429	0.01	0.06	3,607	0.11	0.19	942	0.22	0.27	2,839	0.64	0.33	2,250	0.96	0.10	1,915	0.99	0.03
3rd quintile	3,546	0.02	0.08	3,721	0.14	0.22	1,001	0.31	0.30	3,017	0.72	0.30	2,452	0.98	0.07	1,991	1.00	0.02
4th 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	0.06	2,308	1.00	0.01
5th quintile (highest)	3,893	0.09	0.19	4,161	0.30	0.32	1,158	0.52	0.35	3,644	0.86	0.23	3,163	0.99	0.03	2,534	1.00	0.01
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	0.33	11,670	0.96	0.10	9,193	0.99	0.03
Private school	3,934	0.07	0.17	4,388	0.25	0.30	1,043	0.44	0.33	3,286	0.82	0.25	2,631	0.99	0.04	2,052	1.00	0.01

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A33. Probability of proficiency, mathematics level 5: multiply/divide (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total sample	18,635	0.00	0.03	19,647	0.01	0.07	5,226	0.04	0.13	16,641	0.22	0.29	14,349	0.75	0.32	11,266	0.92	0.20
Sex																		
Male	9,479	0.00	0.04	10,041	0.02	0.09	2,644	0.05	0.15	8,506	0.24	0.31	7,277	0.77	0.32	5,672	0.92	0.19
Female	9,156	0.00	0.01	9,606	0.01	0.05	2,582	0.03	0.11	8,135	0.20	0.27	7,072	0.74	0.32	5,594	0.91	0.20
Race/ethnicity																		
White, Non-Hispanic	10,433	0.00	0.03	11,071	0.02	0.08	2,935	0.06	0.16	9,436	0.29	0.31	8,116	0.83	0.27	6,467	0.95	0.15
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	0.36	1,275	0.84	0.26
Hispanic, race specified	1,588	0.00	0.02	1,624	0.01	0.04	389	0.02	0.07	1,354	0.16	0.25	1,260	0.69	0.33	1,022	0.91	0.20
Hispanic, race not specified	1,800	0.00	0.00	1,834	0.00	0.02	486	0.01	0.03	1,518	0.10	0.19	1,324	0.64	0.35	1,080	0.88	0.22
Asian	897	0.01	0.05	1,088	0.03	0.12	256	0.09	0.22	1,042	0.26	0.32	956	0.80	0.30	785	0.95	0.17
Hawaiian, Other Pacific Islander	187	0.00	0.00	202	0.01	0.05	93	0.01	0.04	188	0.10	0.18	172	0.71	0.33	144	0.92	0.15
American Indian, Alaska Native	354	0.00	0.00	345	0.00	0.04	126	0.01	0.05	298	0.10	0.18	250	0.56	0.36	208	0.79	0.27
More than one race, Non-Hispanic	473	0.00	0.04	472	0.01	0.07	151	0.03	0.11	397	0.23	0.29	380	0.78	0.31	269	0.93	0.18
Socioeconomic status																		
1st quintile (lowest)	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	0.36	1,708	0.80	0.28
2nd quintile	3,429	0.00	0.00	3,607	0.01	0.04	942	0.02	0.09	2,839	0.15	0.23	2,250	0.70	0.33	1,915	0.91	0.20
3rd quintile	3,546	0.00	0.01	3,721	0.01	0.05	1,001	0.03	0.10	3,017	0.21	0.27	2,452	0.79	0.29	1,991	0.95	0.13
4th quintile	3,676	0.00	0.02	3,921	0.02	0.07	1,023	0.04	0.13	3,178	0.28	0.30	2,693	0.85	0.25	2,308	0.96	0.14
5th quintile (highest)	3,893	0.01	0.06	4,161	0.04	0.13	1,158	0.11	0.22	3,644	0.40	0.35	3,163	0.92	0.18	2,534	0.98	0.08
School type																		
Public school	14,701	0.00	0.02	15,259	0.01	0.06	3,971	0.04	0.12	13,292	0.20	0.28	11,670	0.74	0.33	9,193	0.91	0.20
Private school	3,934	0.01	0.04	4,388	0.03	0.11	1,043	0.08	0.18	3,286	0.32	0.32	2,631	0.84	0.25	2,052	0.96	0.13

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A34. Probability of proficiency, mathematics level 6: place value (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	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	0.41	0.39	11,266	0.72	0.37
Sex																		
Male	9,479	0.00	0.01	10,041	0.00	0.02	2,644	0.00	0.04	8,506	0.04	0.12	7,277	0.44	0.40	5,672	0.75	0.36
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	0.37	0.38	5,594	0.70	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	0.50	0.39	6,467	0.80	0.33
Black, Non-Hispanic	2855	0.00	0.00	2,962	0.00	0.01	781	0.00	0.01	2,371	0.01	0.04	1,871	0.20	0.31	1,275	0.53	0.40
Hispanic, race specified	1,588	0.00	0.00	1,624	0.00	0.00	389	0.00	0.00	1,354	0.02	0.07	1,260	0.32	0.37	1,022	0.67	0.38
Hispanic, race not specified	1,800	0.00	0.00	1,834	0.00	0.00	486	0.00	0.00	1,518	0.01	0.04	1,324	0.25	0.34	1,080	0.63	0.40
Asian	897	0.00	0.00	1,088	0.00	0.04	256	0.01	0.06	1,042	0.05	0.16	956	0.51	0.41	785	0.82	0.32
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	0.28	0.32	144	0.66	0.40
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	0.20	0.31	208	0.47	0.42
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	0.43	0.39	269	0.74	0.34
Socioeconomic status																		
1st quintile (lowest)	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	0.18	0.29	1,708	0.47	0.42
2nd 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	0.31	0.36	1,915	0.67	0.37
3rd 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	0.41	0.38	1,991	0.76	0.33
4th quintile	3,676	0.00	0.00	3,921	0.00	0.01	1,023	0.00	0.02	3,178	0.03	0.11	2,693	0.53	0.39	2,308	0.84	0.29
5th quintile (highest)	3,893	0.00	0.01	4,161	0.00	0.03	1,158	0.01	0.06	3,644	0.08	0.18	3,163	0.66	0.37	2,534	0.91	0.22
School type																		
Public school	14,701	0.00	0.01	15,259	0.00	0.01	3,971	0.00	0.03	13,292	0.03	0.10	11,670	0.39	0.39	9,193	0.71	0.38
Private school	3,934	0.00	0.00	4,388	0.00	0.02	1,043	0.01	0.04	3,286	0.05	0.13	2,631	0.50	0.39	2,052	0.84	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A35. Probability of proficiency, mathematics level 7: rate and measurement (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	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	0.13	0.23	11,266	0.42	0.39
Sex																		
Male	9,479	0.00	0.00	10,041	0.00	0.00	2,644	0.00	0.00	8,506	0.00	0.02	7,277	0.15	0.25	5,672	0.45	0.40
Female	9,156	0.00	0.00	9,606	0.00	0.00	2,582	0.00	0.00	8,135	0.00	0.01	7,072	0.10	0.20	5,594	0.39	0.39
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	0.02	8,116	0.17	0.25	6,467	0.52	0.39
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	0.14	1,275	0.19	0.29
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	0.09	0.19	1,022	0.35	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	0.00	1,324	0.06	0.15	1,080	0.30	0.35
Asian	897	0.00	0.00	1,088	0.00	0.00	256	0.00	0.01	1,042	0.01	0.02	956	0.20	0.28	785	0.58	0.40
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	0.14	144	0.38	0.39
American Indian, Alaska Native	354	0.00	0.00	345	0.00	0.00	126	0.00	0.00	298	0.00	0.01	250	0.04	0.11	208	0.20	0.32
More than one race, Non-Hispanic	473	0.00	0.00	472	0.00	0.00	151	0.00	0.00	397	0.00	0.01	380	0.13	0.23	269	0.44	0.41
Socioeconomic status																		
1st quintile (lowest)	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	0.11	1,708	0.19	0.30
2nd 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	0.07	0.16	1,915	0.31	0.34
3rd 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	0.10	0.19	1,991	0.42	0.38
4th 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	0.17	0.25	2,308	0.54	0.38
5th quintile (highest)	3,893	0.00	0.00	4,161	0.00	0.00	1,158	0.00	0.00	3,644	0.01	0.03	3,163	0.28	0.31	2,534	0.70	0.35
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	0.12	0.22	9,193	0.41	0.39
Private school	3,934	0.00	0.00	4,388	0.00	0.00	1,043	0.00	0.00	3,286	0.00	0.02	2,631	0.17	0.26	2,052	0.54	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A36. Probability of proficiency, mathematics level 8: fractions (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N	Mean	SD	N	Mean	SD	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	0.00	14,349	0.01	0.06	11,266	0.13	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	0.00	7,277	0.01	0.08	5,672	0.15	0.31
Female	9,156	0.00	0.00	9,606	0.00	0.00	2,582	0.00	0.00	8,135	0.00	0.00	7,072	0.00	0.04	5,594	0.10	0.25
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	0.00	8,116	0.01	0.08	6,467	0.17	0.32
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	0.00	0.03	1,275	0.02	0.12
Hispanic, race specified	1,588	0.00	0.00	1,624	0.00	0.00	389	0.00	0.00	1,354	0.00	0.00	1,260	0.00	0.04	1,022	0.08	0.22
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	0.00	0.01	1,080	0.06	0.20
Asian	897	0.00	0.00	1,088	0.00	0.00	256	0.00	0.00	1,042	0.00	0.00	956	0.01	0.07	785	0.26	0.38
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	0.00	0.01	144	0.08	0.22
American Indian, Alaska Native	354	0.00	0.00	345	0.00	0.00	126	0.00	0.00	298	0.00	0.00	250	0.00	0.01	208	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	0.01	0.04	269	0.18	0.33
Socioeconomic status																		
1st quintile (lowest)	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	0.00	0.01	1,708	0.04	0.15
2nd quintile	3,429	0.00	0.00	3,607	0.00	0.00	942	0.00	0.00	2,839	0.00	0.00	2,250	0.00	0.03	1,915	0.05	0.18
3rd quintile	3,546	0.00	0.00	3,721	0.00	0.00	1,001	0.00	0.00	3,017	0.00	0.00	2,452	0.00	0.04	1,991	0.10	0.24
4th quintile	3,676	0.00	0.00	3,921	0.00	0.00	1,023	0.00	0.00	3,178	0.00	0.00	2,693	0.01	0.09	2,308	0.16	0.31
5th quintile (highest)	3,893	0.00	0.00	4,161	0.00	0.00	1,158	0.00	0.00	3,644	0.00	0.00	3,163	0.02	0.10	2,534	0.32	0.39
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	0.00	11,670	0.01	0.06	9,193	0.12	0.28
Private school	3,934	0.00	0.00	4,388	0.00	0.00	1,043	0.00	0.00	3,286	0.00	0.00	2,631	0.01	0.07	2,052	0.18	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, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A37. Probability of proficiency, mathematics level 9: area and volume (range of possible values: 0.0 to 1.0): School years 1998–99, 1999–2000, 2001–02, and 2003–04

Characteristic	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
	N[1]	Mean	SD[2]	N	Mean	SD	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	0.00	14,349	0.00	0.01	11,266	0.02	0.07
Sex																		
Male	9,479	0.00	0.00	10,041	0.00	0.00	2,644	0.00	0.00	8,506	0.00	0.00	7,277	0.00	0.01	5,672	0.02	0.08
Female	9,156	0.00	0.00	9,606	0.00	0.00	2,582	0.00	0.00	8,135	0.00	0.00	7,072	0.00	0.00	5,594	0.01	0.05
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	0.00	8,116	0.00	0.01	6,467	0.02	0.08
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	0.00	0.00	1,275	0.00	0.02
Hispanic, race specified	1,588	0.00	0.00	1,624	0.00	0.00	389	0.00	0.00	1,354	0.00	0.00	1,260	0.00	0.01	1,022	0.01	0.05
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	0.00	0.00	1,080	0.01	0.03
Asian	897	0.00	0.00	1,088	0.00	0.00	256	0.00	0.00	1,042	0.00	0.00	956	0.00	0.00	785	0.05	0.12
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	0.00	0.00	144	0.01	0.03
American Indian, Alaska Native	354	0.00	0.00	345	0.00	0.00	126	0.00	0.00	298	0.00	0.00	250	0.00	0.00	208	0.00	0.02
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	0.00	0.00	269	0.02	0.05
Socioeconomic status																		
1st quintile (lowest)	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	0.00	0.00	1,708	0.00	0.03
2nd quintile	3,429	0.00	0.00	3,607	0.00	0.00	942	0.00	0.00	2,839	0.00	0.00	2,250	0.00	0.00	1,915	0.01	0.04
3rd quintile	3,546	0.00	0.00	3,721	0.00	0.00	1,001	0.00	0.00	3,017	0.00	0.00	2,452	0.00	0.00	1,991	0.01	0.04
4th quintile	3,676	0.00	0.00	3,921	0.00	0.00	1,023	0.00	0.00	3,178	0.00	0.00	2,693	0.00	0.01	2,308	0.02	0.06
5th quintile (highest)	3,893	0.00	0.00	4,161	0.00	0.00	1,158	0.00	0.00	3,644	0.00	0.00	3,163	0.00	0.01	2,534	0.05	0.11
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	0.00	11,670	0.00	0.01	9,193	0.02	0.07
Private school	3,934	0.00	0.00	4,388	0.00	0.00	1,043	0.00	0.00	3,286	0.00	0.00	2,631	0.00	0.01	2,052	0.02	0.07

[1] Number in sample.
[2] Standard deviation.
NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A38. Percent of children at or above modal reading proficiency for each grade: School years 1998–99, 1999–2000, 2001–02, and 2003–04

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		Round 6 Modal Level=7	
	N[1]	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Total sample	16,739	64.50	17,691	48.60	4,740	65.20	15,226	77.60	13,259	71.40	10,583	71.10
Sex												
Male	8,536	60.80	9,003	45.20	2,394	60.50	7,736	73.40	6,738	68.30	5,290	68.30
Female	8,203	68.50	8,688	52.10	2,346	70.10	7,490	82.10	6,521	74.80	5,293	74.00
Race/ethnicity												
White, Non-Hispanic	9,887	69.80	10,402	55.00	2,787	72.50	8,931	83.00	7,510	80.00	6,124	79.30
Black, Non-Hispanic	2,744	59.90	2,735	33.90	714	51.70	2,129	67.60	1,718	58.70	1,187	58.20
Hispanic, race specified	1,126	51.80	1,225	44.30	295	63.00	1,142	73.90	1,160	65.00	951	62.20
Hispanic, race not specified	1,130	46.70	1,319	38.00	343	50.30	1,200	65.00	1,224	53.10	1,003	52.90
Asian	845	81.50	1,017	63.40	240	68.90	970	84.90	905	73.80	740	76.20
Hawaiian, Other Pacific Islander	179	64.00	185	35.20	90	42.00	167	75.80	154	58.30	132	61.30
American Indian, Alaska Native	336	34.90	324	26.80	121	24.80	276	50.00	219	40.30	187	42.50
More than one race, Non-Hispanic	447	62.70	438	44.60	142	65.30	378	82.30	349	73.90	244	81.50
Socioeconomic status												
1st quintile (lowest)	2,514	41.70	2,726	27.00	688	40.10	2,110	59.00	1,844	49.10	1,565	44.00
2nd quintile	3,114	56.60	3,259	39.10	861	55.70	2,577	73.80	2,065	65.70	1,813	64.80
3rd quintile	3,294	64.40	3,416	48.30	935	67.70	2,798	80.20	2,253	75.10	1,879	75.90
4th quintile	3,462	73.70	3,643	58.30	964	76.50	3,016	85.80	2,479	81.70	2,170	83.70
5th quintile (highest)	3,629	83.80	3,924	70.10	1,100	83.40	3,511	90.60	2,969	91.10	2,410	90.90
School type												
Public school	13,073	61.20	13,614	45.30	3,565	63.10	12,054	75.90	10,749	69.70	8,610	69.60
Private school	3,666	82.80	4,077	66.20	986	83.40	3,125	89.20	2,463	84.80	1,951	82.70

[1] Number in sample.

NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table A39. Percent of children at or above modal mathematics proficiency for each grade: School years 1998–99, 1999–2000, 2001–02, and 2003–04

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		Round 6 Modal Level=7	
	N[1]	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Total sample	18,149	55.40	18,945	53.40	5,060	70.70	16,133	70.20	13,998	74.30	10,874	73.20
Sex												
Male	9,178	52.80	9,641	52.60	2,550	69.90	8,199	70.20	7,081	75.70	5,479	77.50
Female	8,971	58.10	9,304	54.20	2,510	71.60	7,934	70.20	6,917	72.80	5,395	68.70
Race/ethnicity												
White, Non-Hispanic	10,104	66.10	10,703	64.90	2,848	80.40	9,223	78.10	7,934	82.50	6,244	81.10
Black, Non-Hispanic	2,800	44.50	2,838	37.20	753	60.00	2,287	57.90	1,837	53.60	1,229	53.70
Hispanic, race specified	1,552	40.50	1,555	38.60	374	61.50	1,293	62.20	1,223	68.10	977	67.50
Hispanic, race not specified	1,777	27.70	1,766	30.10	464	46.40	1,426	55.70	1,283	66.40	1,047	65.00
Asian	873	69.50	1,049	61.40	252	72.30	1,016	70.40	922	77.30	759	84.50
Hawaiian, Other Pacific Islander	182	53.40	196	39.20	91	56.40	179	54.10	167	70.70	140	82.50
American Indian, Alaska Native	351	35.00	331	34.20	123	41.90	287	48.20	239	56.90	203	54.30
More than one race, Non-Hispanic	463	57.90	459	51.90	146	70.70	385	69.60	373	80.50	259	72.00
Socioeconomic status												
1st quintile (lowest)	3,212	28.70	3,278	27.00	864	42.40	2,434	50.10	1,944	54.60	1,647	47.50
2nd quintile	3,353	47.40	3,485	46.30	911	64.50	2,737	64.40	2,188	69.50	1,844	67.50
3rd quintile	3,458	58.00	3,593	55.90	959	79.20	2,942	72.70	2,405	78.00	1,932	78.30
4th quintile	3,568	67.50	3,778	65.20	995	81.50	3,116	79.60	2,622	83.60	2,224	85.20
5th quintile (highest)	3,751	78.80	4,031	76.70	1,130	88.30	3,566	86.80	3,105	90.40	2,446	91.30
School type												
Public school	14,332	52.30	14,692	50.40	3,831	68.20	12,856	68.70	11,390	73.40	8,866	71.70
Private school	3,817	73.50	4,253	70.00	1,023	89.80	3215	81.80	2,562	82.50	1,989	85.00

[1] Number in sample.

NOTE: Table estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Estimates for kindergarten through fifth 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), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

APPENDIX B

ECLS-K ITEM PARAMETERS BY ROUNDS

Table B1. Reading assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, and 2003–04

Item label	IRT parameters			Used in	Mean and standard deviation of theta[4]					
	a[1]	b[2]	c[3]	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
CANDLE	0.78	-3.45	0.15	K1						
POURINT	0.85	-2.70	0.14	K1						
CEREAL	1.13	-2.65	0.00	K1						
DECORATD	0.75	-2.57	0.13	K1						
BEGBIKE	1.62	-1.83	0.00	K1	*					
BEGIN	0.89	-1.75	0.00	K1	*					
VEGETBLE	0.71	-1.59	0.11	K1	*	*				
LETRECD	2.66	-1.58	0.00	K1	*	*				
LETRECF	3.02	-1.54	0.00	K1	*	*				
LETRECM	2.66	-1.53	0.00	K1	*	*				
LETRECT	2.83	-1.46	0.00	K1	*	*				
COULDNOT	0.88	-1.40	0.00	K1	*	*	*			
KAYLAFLY	0.65	-1.34	0.00	K1	*	*	*			
NEXTLINE	1.10	-1.31	0.00	K1	*	*	*			
STORYEND	1.27	-1.30	0.00	K1	*	*	*			
TIME	0.99	-1.30	0.00	K1	*	*	*			
TRUNK	0.71	-1.25	0.00	K1	-1.20	*	*			
BEGP	1.72	-1.12	0.00	K1	(.50)	*	*			
BEGR	2.30	-1.10	0.00	K1	*	*	*			
BEGL	2.27	-1.06	0.00	K1	*	*	*			
AWARDING	0.96	-0.97	0.27	K1	*	*	*			
JOGGING	1.19	-0.94	0.10	K1	*	*	*			
COULD	0.59	-0.91	0.00	K1	*	*	*			
ENDL	2.14	-0.88	0.00	K1	*	*	*			
MOM	2.31	-0.88	0.00	K1	*	*	*			
ENDF	1.78	-0.84	0.00	K1	*	*	*			
YELLOW	1.88	-0.78	0.00	K1	*	*	*			
BEGB	1.41	-0.76	0.00	K1	*	*	*			
BEGWORD	0.85	-0.73	0.00	K1	*	*	*	*		
ENDP	1.61	-0.70	0.00	K1	*	*	*	*		
QMARK	1.10	-0.69	0.00	K1	*	-0.63	*	*		
ENDD	1.66	-0.56	0.00	K1	*	(.50)	*	*		
YOU	2.69	-0.54	0.00	K1	*	*	*	*		
ORPIG	2.07	-0.43	0.00	K1	*	*	*	*		
ORSAT	2.72	-0.40	0.00	K1	*	*	-.39	*		
ORTAIL	3.06	-0.30	0.00	K1	*	*	(.51)	*		
RUNS	3.33	-0.26	0.00	K1,3	*	*	*	*		
ORHAND	3.14	-0.23	0.00	K1	*	*	*	*		
NEEDHOME-	4.00	-0.18	0.13	K1	*	*	*	*		

See notes at end of table.

www.manaraa.com

Table B1. Reading assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

Item label	IRT parameters			Used in grades	Mean and standard deviation of theta[4]					
	a[1]	b[2]	c[3]		Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
WENT	3.21	-0.16	0.00	K1,3		*	*	*		
DOWN	3.92	-0.12	0.00	K1,3		*	*	*		
BOYBIRD	3.63	-0.11	0.19	K1		*	*	*		
JEEP	3.03	-0.10	0.00	K1,3		*	*	*		
GIRLRED	1.54	-0.07	0.00	K1		*	*	*		
FISHING	4.86	-0.07	0.00	K1		*	*	*		
CANINBAG	2.05	-0.06	0.21	K1		*	*	*		
KITNBED	3.09	-0.05	0.16	K1		*	*	*		
CATCH	3.67	-0.03	0.00	K1		*	*	*		
MAKE	1.24	0.01	0.15	1		*	*	*		
KNOW	2.53	0.11	0.00	K1		*	*	*		
LIGHT	4.00	0.12	0.00	K1		*	*	*		
KIMCAD	4.76	0.13	0.50	K1		*	*	*		
ELEPHANT	3.67	0.14	0.00	K1		*	*	*		
BACKPACK	2.84	0.22	0.14	K1,3,5		*	*	.22		
LIKEDRY	4.26	0.26	0.25	K1		*	*	(.48)	*	
FLATTIRE	3.36	0.26	0.15	K1		*	*	*	*	
LISTEN	3.64	0.29	0.11	K1,3,5		*	*	*	*	
WRONG	3.50	0.30	0.00	K1		*	*	*	*	
RIDEBIKE	3.40	0.36	0.18	K1,3,5		*	*	*	*	
SIZES	4.18	0.40	0.13	K1,3,5			*	*	*	
CHOCCAKE	4.93	0.41	0.17	K1			*	*	*	
QUIET	3.30	0.50	0.00	K1,3			*	*	*	
RDBIGKY	1.76	0.50	0.00	3			*	*	*	
DOGHOUSE	2.64	0.51	0.16	K1			*	*	*	
ENVELOPE	3.54	0.52	0.00	K1			*	*	*	
RDFINGRY	2.25	0.54	0.10	3			*	*	*	*
RDLETR	2.84	0.56	0.27	3,5			*	*	*	*
THROUGH	2.56	0.58	0.00	K1,3,5			*	*	*	*
RDMARIAB	1.56	0.59	0.28	3,5			*	*	*	*
RDGROSR	2.13	0.59	0.19	3,5			*	*	*	*
RDLIKE	1.33	0.63	0.09	3,5			*	*	*	*
RDDANGRY	1.49	0.63	0.07	3			*	*	*	*
RDTIME	3.23	0.64	0.32	3,5				*	*	*
RDENDR	2.96	0.65	0.00	3,5				*	*	*
RAGE	3.33	0.66	0.00	K1,3				*	*	*
MARCHED	4.47	0.67	0.20	K1				*	*	*
RDFEELSR	3.72	0.67	0.17	3,5				*	*	*
CATNAME	2.50	0.68	0.00	1				*	*	*
WTLESS	5.21	0.69	0.00	K1,3,5				*	*	*
RDSAMER	2.45	0.71	0.00	3,5				*	*	*
RDBEARY	2.42	0.71	0.08	3				*	*	*

See notes at end of table.

www.manaraa.com

Table B1. Reading assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

Item label	IRT parameters			Used in	Mean and standard deviation of theta[4]					
	a[1]	b[2]	c[3]	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
TOIL	2.38	0.71	0.00	K1,3				*	*	*
CORNER	2.48	0.73	0.00	K1,3				*	*	*
RDGEORGR	3.22	0.74	0.24	3,5				*	*	*
OWNRNAME	2.49	0.75	0.00	1				*	*	*
REQUIRE	4.22	0.76	0.00	K1,3				*	*	*
RDTANZAR	3.14	0.77	0.00	3,5				*	*	*
CAPTURE	2.79	0.78	0.00	K1,3				*	*	*
RDFACTY	2.98	0.78	0.17	3				*	*	*
WEB	1.94	0.80	0.00	K1,3				*	*	*
RDDOCR	2.56	0.81	0.00	3,5				*	*	*
RDKINDY	1.75	0.81	0.10	3				*	*	*
UNUSUAL	4.61	0.82	0.00	K1				*	*	*
RDBSITY	3.46	0.83	0.00	3				*	*	*
MOISTURE	3.44	0.83	0.00	K1,3,5				*	*	*
RDSISR	2.86	0.83	0.14	3,5				*	*	*
MOTHER	1.23	0.84	0.00	5				*	*	*
RDTRUEY	2.34	0.85	0.11	3				*	*	*
RDSTORY	2.33	0.86	0.18	3,5				*	*	*
RECIPE	3.62	0.87	0.19	K1				*	*	*
RDSTRAGY	1.74	0.88	0.00	3				*	*	*
MAINPROB	1.46	0.90	0.14	5				*	*	*
PREDICT	3.07	0.90	0.12	5				*	*	*
RDWAY	1.82	0.90	0.00	3,5				*	*	*
RDKNIGHT	2.40	0.93	0.00	3,5				*	*	*
INGREDNT	4.72	0.94	0.19	K1				*	*	*
RDJAMEDR	1.88	0.94	0.00	3,5				*	*	*
EXAMPLE	2.72	0.95	0.14	5				*	*	*
RDCLUER	2.80	0.96	0.00	3,5				*	.96	*
RDBOWY	2.77	0.97	0.13	3,5				*	(.36)	*
RDTRAINY	3.57	0.98	0.13	3,5				*	*	*
RDSUPRIR	2.64	0.99	0.00	3,5				*	*	*
MOREINFO	1.84	1.03	0.00	1				*	*	*
MYSTERLY	3.50	1.03	0.00	K1				*	*	*
WHYNO	1.26	1.04	0.00	1				*	*	*
IMP_UNDR	1.68	1.04	0.11	5				*	*	*
DR_ROSE	2.74	1.05	0.12	5				*	*	*
RDFRICTY	1.99	1.06	0.00	3				*	*	*
APPROX	2.44	1.06	0.00	1				*	*	*
RDTEARB	2.66	1.10	0.26	3,5				*	*	*
WAGES	2.92	1.10	0.00	K1,5				*	*	*
RDSAFER	2.57	1.11	0.00	3,5				*	*	*
MAINIDEA	1.90	1.12	0.29	1				*	*	*
VICIOUS	3.65	1.13	0.00	K1				*	*	*

See notes at end of table.

www.manaraa.com

Table B1. Reading assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

Item label	IRT parameters			Used in	Mean and standard deviation of theta[4]					
	a[1]	b[2]	c[3]	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
RDBAKEDB	1.55	1.14	0.13	3,5				*	*	*
SLUDGE	2.38	1.17	0.10	5				*	*	*
RDPOUCHY	2.82	1.17	0.04	3				*	*	*
RDTHREEB	2.02	1.17	0.00	3,5				*	*	*
RDMOVEBY	1.27	1.18	0.05	3,5				*	*	*
CORNERS4	2.17	1.18	0.00	5				*	*	*
RDMALEBY	2.30	1.19	0.03	3					*	*
DIFFRNT	2.45	1.20	0.09	5					*	*
RDLIKER	2.52	1.21	0.00	3,5					*	*
RDDOMEST	2.43	1.21	0.00	3,5					*	*
RDAPOSTY	1.76	1.21	0.00	3					*	*
SPRING	3.81	1.23	0.13	5					*	*
RDBABONY	1.54	1.24	0.07	3					*	*
STRANDS	1.73	1.24	0.00	K1,3					*	*
SLOW_LRN	2.42	1.24	0.00	5					*	*
COMPASS	3.50	1.24	0.10	5					*	*
RDDIFFR	1.55	1.24	0.00	3,5					*	*
RDINFLUB	2.22	1.24	0.00	3,5					*	*
ABOUT	2.39	1.24	0.08	5					*	*
RDPROBLY	1.55	1.25	0.04	3,5					*	*
CRITCISM	3.16	1.25	0.00	K1,3,5					*	1.26
OVATIONS	1.82	1.27	0.24	5					*	(.36)
RDBRETY	2.43	1.27	0.24	3,5					*	*
DEPART	3.63	1.30	0.12	5					*	*
PREFRNCE	1.73	1.31	0.00	K1,3,5					*	*
WHY_LEFT	3.26	1.32	0.00	5					*	*
RDJOSHB	1.49	1.32	0.00	3,5					*	*
RDRACHLB	1.94	1.33	0.13	3,5					*	*
RDTHEMEB	2.19	1.34	0.07	3,5					*	*
WHYCONTR	2.07	1.39	0.06	5					*	*
DESCRIBE	2.69	1.40	0.12	1					*	*
RDMICROB	2.33	1.41	0.00	3,5					*	*
AMBITIO	2.50	1.41	0.00	K1,3					*	*
ON_MESA	2.23	1.42	0.00	5					*	*
RDSOLVEY	2.21	1.43	0.05	3,5					*	*
ALIGNMNT	2.19	1.45	0.00	K1,5					*	*
RDPERSNB	2.19	1.47	0.00	3,5					*	*
MTPCOMP	2.53	1.47	0.00	5					*	*
SUMMARY	1.26	1.47	0.10	5					*	*
RDHELPY	1.82	1.49	0.03	3,5					*	*
RDCOMPRB	1.88	1.50	0.00	3,5					*	*
LIKE_DIS	1.36	1.52	0.00	5					*	*

See notes at end of table.

www.manaraa.com

Table B1. Reading assessment IRT item parameters: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

Item label	IRT parameters			Used in grades	Mean and standard deviation of theta[4]					
	a[1]	b[2]	c[3]		Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
ERUPT2	2.17	1.53	0.00	5					*	*
SUPPORT	1.68	1.54	0.10	5					*	*
AUTHOR	1.50	1.57	0.00	5					*	*
PSYCHLG	1.43	1.63	0.00	5					*	*
RDGUESS	1.31	1.67	0.11	3,5					*	*
RDHOAXB	3.16	1.68	0.00	3,5					*	*
DOUBT1	4.74	1.68	0.00	5					*	*
RDCROPB	3.20	1.68	0.17	3,5					*	*
ADVANCES	1.13	1.69	0.00	5						*
INSUFFIC	1.97	1.71	0.00	5						*
DOUBT2	4.71	1.74	0.00	5						*
DCIRCLB	1.91	1.77	0.06	3,5						*
TONE	1.89	1.84	0.09	5						*
RDVORTXB	3.60	1.85	0.23	3,5						*
MAINPURP	1.78	1.86	0.08	5						*
THEORY2	1.50	1.86	0.00	5						*
RDWAGON	2.49	1.96	0.21	3,5						*
BELLGRNT	0.65	2.08	0.00	5						
RDANOMAB	0.59	2.83	0.00	3						
RDEMBOLY	0.98	2.91	0.00	3						
PROFLEV1	3.50	-1.46	0.00	K1						
PROFLEV2	3.22	-0.90	0.00	K1						
PROFLEV3	3.05	-0.61	0.00	K1						
PROFLEV4	4.25	-0.08	0.00	K1,3						
PROFLEV5	3.00	0.31	0.00	K1,3,5						
PROFLEV6	3.50	0.77	0.00	3,5						
PROFLEV7	5.93	1.06	0.00	3,5						
PROFLEV8	2.45	1.35	0.00	3,5						
PROFLEV9	6.13	1.87	0.00	5						

[1] Parameter for discrimination.
[2] Parameter for difficulty.
[3] Parameter for guessing.
[4] Mean and standard deviation of theta ability estimate

NOTE: Item responses from kindergarten through fifth grade were pooled for IRT calibration to produce parameter estimates on a common scale. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). The grades in which items appeared on assessment forms are noted. Mean and standard deviation of theta ability estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Asterisks mark the range corresponding to 2 standard deviations below and above the mean ability for the round.

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

www.manaraa.com

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

Item label	IRT parameters			Used in	Mean and standard deviation of theta[4]					
	a[1]	b[2]	c[3]	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
2CRAYONS	1.77	-2.87	0.00	K1						
3BANANAS	0.89	-2.61	0.08	K1						
SQUARE	1.05	-2.44	0.13	K1						
NUMBER 4	3.53	-1.89	0.00	K1	*					
# STRAW	1.21	-1.79	0.00	K1	*					
STICKBAT	1.01	-1.67	0.05	K1	*					
3-1PENCL	0.91	-1.67	0.00	K1	*					
NUMBER 7	3.12	-1.64	0.00	K1	*					
#VANILLA	1.35	-1.58	0.00	K1	*	*				
#CHOC	1.43	-1.41	0.00	K1	*	*				
NUMBER 9	2.65	-1.40	0.00	K1	*	*				
PNTBRUSH	1.71	-1.32	0.21	K1	*	*	*			
COUNT 20	1.28	-1.28	0.00	K1	*	*	*			
4LINES	0.64	-1.27	0.16	K1	*	*	*			
6BANANAS	1.24	-1.18	0.00	K1	-1.14	*	*			
LG-SM-SM	1.66	-1.08	0.28	K1	(.50)	*	*			
SM-LG-SM	1.49	-1.08	0.23	K1	*	*	*			
NUMBER17	2.14	-0.98	0.00	K1	*	*	*			
000X	1.19	-0.92	0.19	K1	*	*	*			
NUMBER23	2.14	-0.83	0.00	K1	*	*	*			
3RD LINE	2.06	-0.80	0.00	K1	*	*	*			
3+2 CARS	1.43	-0.79	0.00	K1	*	*	*			
_ 78910	2.00	-0.77	0.00	K1	*	*	*			
HALFOVAL	1.01	-0.77	0.22	K1	*	*	*			
2+3STICK	1.60	-0.72	0.00	K1	*	*	*	*		
#BUGS	1.56	-0.68	0.22	K1	*	*	*	*		
2 + 2	3.00	-0.66	0.00	K1	*	-0.62	*	*		
3 + 3	4.00	-0.56	0.00	K1	*	(.49)	*	*		
1 + 7	1.49	-0.55	0.00	K1	*	*	*	*		
TEAMS_R	1.13	-0.54	0.15	3	*	*	*	*		
VICKS_R	2.42	-0.44	0.00	3	*	*	*	*		
8-6CRAYN	1.35	-0.43	0.00	K1	*	*	*	*		
3 + 4	2.29	-0.34	0.00	K1	*	*	-0.34	*		
5-1ORANG	1.99	-0.31	0.12	K1	*	*	(.50)	*		
2+5MARBL	1.43	-0.30	0.00	K1,3	*	*	*	*		
SHAPES	0.70	-0.26	0.19	K1	*	*	*	*		
PATTERN	1.45	-0.22	0.21	K1	*	*	*	*		
2+5CIRCL	1.69	-0.21	0.00	K1	*	*	*	*		
12 BY 2S	2.08	-0.19	0.00	K1,3	*	*	*	*		
3+7PENNY	2.07	-0.13	0.00	K1,3	*	*	*	*		
51015_25	2.33	-0.04	0.00	K1,3		*	*	*		
ORANGE_R	1.74	-0.03	0.14	3		*	*	*		
11 + 3	2.29	0.00	0.00	K1		*	*	*		

See notes at end of table.

www.manaraa.com

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

Item label	IRT parameters			Used in grades	Mean and standard deviation of theta[4]					
	a[1]	b[2]	c[3]		Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
7-3	2.82	0.03	0.00	K1		*	*	*		
9-2	2.96	0.05	0.00	K1		*	*	*		
PATHS_R	1.16	0.05	0.00	3		*	*	*		
6+7	2.32	0.13	0.00	K1		*	*	*	*	
12 + 6	1.90	0.23	0.00	K1		*	*	.19	*	
# MORE	1.97	0.25	0.00	K1		*	*	(.46)	*	
MOST_Y	2.68	0.26	0.00	3		*	*	*	*	
2-1+2	1.80	0.28	0.00	K1		*	*	*	*	
RULER_R	1.29	0.34	0.00	3		*	*	*	*	
A13_79	1.80	0.35	0.00	K1,3,5		*	*	*	*	
4+4-2	2.26	0.37	0.00	K1,3			*	*	*	
SIDES_R	1.61	0.38	0.12	3			*	*	*	*
PAGES_R	2.35	0.38	0.20	3			*	*	*	*
17 – 4	2.62	0.39	0.00	K1			*	*	*	*
COST_10	2.27	0.40	0.00	K1,3,5			*	*	*	*
12-9	2.57	0.41	0.00	K1			*	*	*	*
26 + 20	2.67	0.47	0.00	K1			*	*	*	*
CARS15_5	2.38	0.47	0.00	K1,3,5			*	*	*	*
FEWEST_Y	2.51	0.51	0.00	3			*	*	*	*
SQUARE_R	0.96	0.52	0.00	3			*	*	*	*
CUBES10	0.93	0.53	0.00	3,5			*	*	*	*
HOWMANY$	1.60	0.59	0.00	K1,3			*	*	*	*
CANDY8_2	2.34	0.60	0.00	K1,3,5			*	*	*	*
BEADS_R	4.06	0.63	0.00	3			*	*	*	*
NEXT78	2.50	0.64	0.00	3,5			*	*	*	*
12-? PEN	2.55	0.64	0.00	K1,3			*	*	*	*
HEADSUP	1.22	0.65	0.00	K1,3			*	*	*	*
24-14BKS	2.77	0.66	0.00	K1			*	*	*	*
MEANS_R	2.83	0.71	0.00	3				*	*	*
EQUAL_R	2.99	0.71	0.17	3				*	*	*
DO_ADD4	2.23	0.71	0.00	3,5				*	*	*
MONEY_R	3.44	0.72	0.00	3				*	*	*
TIME1030	2.02	0.73	0.00	3,5				*	*	*
POINTS_R	1.86	0.75	0.29	3				*	*	*
SCORE_Y	3.14	0.76	0.00	3				*	*	*
GOALS	2.28	0.77	0.00	K1,3				*	*	*
PAPERS	2.81	0.78	0.00	3				*	*	*
NICKELS	2.43	0.80	0.00	3				*	*	
17CENTS	2.83	0.83	0.00	K1				*	*	*
MORE1_Y	3.63	0.85	0.00	3				*	*	*
NUMBER60	3.34	0.86	0.00	3,5				*	*	*
BDCAKE	2.19	0.87	0.00	K1				*	*	*

See notes at end of table.

www.manaraa.com

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

Item label	IRT parameters			Used in	Mean and standard deviation of theta[4]					
	a[1]	b[2]	c[3]	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
CUBESIDE	1.33	0.88	0.00	3,5				*	*	*
FEWER_Y	3.62	0.91	0.00	3				*	.91	*
NEXT120	2.83	0.91	0.00	3,5				*	(.42)	*
CHART_64	1.95	0.93	0.00	3,5				*	*	*
AGEGRAPH	1.60	0.93	0.00	5				*	*	*
BOX_700	3.33	0.95	0.00	3,5				*	*	*
NUMBER	2.48	0.95	0.00	3				*	*	*
SPOONS	2.72	0.96	0.00	3,5				*	*	*
CANDY27	2.91	0.97	0.00	5				*	*	*
TREES100	2.72	0.99	0.00	5				*	*	*
COLORSYM	1.94	0.99	0.00	3,5				*	*	*
FRIES	2.78	0.99	0.00	3				*	*	*
CHILDR_Y	2.60	1.00	0.00	3				*	*	*
STAR-Y	1.32	1.01	0.00	3				*	*	*
PAGES78	2.48	1.02	0.08	3,5				*	*	*
BOXSHELF	1.74	1.03	0.00	5				*	*	*
SECOND_Y	2.45	1.04	0.00	3				*	*	*
A568214K	2.65	1.04	0.00	3,5				*	*	*
A1ST_X5	1.98	1.04	0.22	5				*	*	*
PATTRN18	1.33	1.07	0.00	5				*	*	*
BIKETIME	2.16	1.08	0.00	5				*	*	*
FRUIT	1.88	1.09	0.12	3				*	*	*
24/4 TAB	1.60	1.12	0.00	K1					*	*
SCALE_	1.87	1.16	0.00	5					*	*
CHARGE_5	2.05	1.19	0.00	3,5					*	*
MARIA310	2.56	1.21	0.00	3,5					*	*
CARDS579	2.21	1.22	0.00	3,5					*	*
LEMONS24	2.28	1.25	0.00	5					*	*
TILES	1.49	1.26	0.00	3					*	*
PAIR_100	3.08	1.29	0.13	3,5					*	*
AREA_B	1.70	1.31	0.00	3					*	*
LARGER_B	1.82	1.34	0.00	3					*	1.34
PENCIL_Y	1.18	1.38	0.05	3					*	(.48)
GREW4_	1.85	1.38	0.00	3,5					*	*
LOUISA13	2.67	1.40	0.00	3,5					*	*
EQUAL_B	1.91	1.40	0.08	3					*	*
AGE1_4	2.95	1.42	0.23	5					*	*
STU1_444	3.60	1.44	0.00	5					*	*
GAMESCOR	1.91	1.45	0.11	5					*	*
NUMBE2_B	2.12	1.51	0.00	3					*	*
LONGSTEP	1.73	1.51	0.00	5					*	*
MIN_BLOW	2.07	1.52	0.00	3,5					*	*

See notes at end of table.

www.manaraa.com

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

Item label	IRT parameters			Used in	Mean and standard deviation of theta[4]					
	a[1]	b[2]	c[3]	grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
BEADSWHT	1.87	1.56	0.00	5					*	*
TALL75_	2.51	1.57	0.09	3,5					*	*
CHANGE	1.83	1.58	0.00	K1					*	*
MARBLES	2.69	1.59	0.00	3,5					*	*
BANKER_	1.69	1.62	0.00	3,5					*	*
MYSTER_B	2.52	1.64	0.00	3					*	*
OJ_30OZ	2.27	1.69	0.00	5					*	*
FRAME3FT	2.25	1.70	0.00	5					*	*
MARK_DOT	1.98	1.73	0.00	3,5					*	*
EDGECUBE	1.05	1.75	0.00	3,5					*	*
HOOP2_5	3.43	1.79	0.00	5						*
SAMEFRAC	1.95	1.82	0.00	3,5						*
SHADED_2	2.51	1.86	0.11	5						*
BUDGETFR	2.15	1.87	0.15	5						*
PIZZA	2.52	1.91	0.05	5						*
FRAC3_4	2.74	1.91	0.09	5						*
SALESTAX	1.16	1.93	0.20	5						*
OPOSITIV	2.48	1.94	0.00	5						*
AREAPLAY	2.02	1.98	0.05	5						*
FENCE_B	2.17	2.00	0.00	3						*
DIFF_88	2.28	2.08	0.08	5						*
SHADED_3	3.23	2.08	0.00	5						*
MEASDIAM	2.10	2.23	0.00	5						*
CARPET	2.72	2.41	0.00	5						
PRISMVOL	1.67	2.44	0.00	5						
TILESCOV	1.83	2.65	0.00	3,5						
PROFLEV1	3.55	-1.93	0.00	K1						
PROFLEV2	3.04	-1.19	0.00	K1						
PROFLEV3	4.30	-0.65	0.00	K1						
PROFLEV4	3.61	-0.04	0.00	K1,3						
PROFLEV5	4.40	0.58	0.00	K1,3,5						
PROFLEV6	5.90	1.03	0.00	3,5						
PROFLEV7	4.68	1.45	0.00	3,5						
PROFLEV8	8.32	1.90	0.00	5						
PROFLEV9	4.24	2.43	0.00	5						

[1] Parameter for discrimination.
[2] Parameter for difficulty.
[3] Parameter for guessing.
[4] Mean and standard deviation of theta ability estimate

NOTE: Item responses from kindergarten through fifth grade were pooled for IRT calibration to produce parameter estimates on a common scale. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). The grades in which items appeared on assessment forms are noted. Mean and standard deviation of theta ability estimates are based on cross-sectional weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0). Asterisks mark the range corresponding to 2 standard deviations below and above the mean ability for the round.

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

www.manaraa.com

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

Item label	IRT parameters			Used in grades	Mean and standard deviation of theta[4]	
	a[1]	b[2]	c[3]		Round 5	Round 6
RBULB	0.72	-2.42	0.13	3		
RENRGY	1.03	-2.38	0.15	3		
RPLANT	1.18	-2.30	0.15	3		
RORGAN	0.46	-2.03	0.13	3	*	
RTOOL	0.73	-1.94	0.11	3	*	
ROUIMM	0.81	-1.92	0.03	3,5	*	
RDSAST	1.20	-1.83	0.08	3	*	
RFGRPS	0.48	-1.80	0.20	3	*	
RFORMS	0.76	-1.70	0.15	3	*	
YPLAIN	0.62	-1.55	0.00	3	*	
RWINGS	1.31	-1.45	0.08	3,5	*	*
RANIML	0.76	-1.38	0.09	3	*	*
ROUFRZ	1.17	-1.26	0.09	3,5	*	*
ROCCUR	1.26	-1.15	0.21	3	*	*
WHCHPREY	0.98	-1.09	0.00	5	*	*
RSEEDS	0.73	-1.04	0.07	3	*	*
ROUTAP	0.48	-0.96	0.01	3,5	*	*
ROUJUN	1.12	-0.95	0.00	3,5	*	*
RTHING	0.94	-0.93	0.21	3	*	*
RWATER	0.88	-0.91	0.10	3	*	*
YDSAST	0.64	-0.90	0.13	3	*	*
RSUNIS	0.97	-0.86	0.28	3	*	*
ROUERT	0.71	-0.81	0.12	3,5	*	*
ROUBRN	1.12	-0.79	0.00	3,5	*	*
RFISHB	0.83	-0.76	0.10	3	*	*
RSHAPE	0.89	-0.75	0.18	3	*	*
RHEART	1.06	-0.71	0.00	3,5	*	*
RPWDER	0.81	-0.70	0.15	3	*	*
ROUJAR	0.49	-0.67	0.00	3,5	*	*
CUTSCAB	0.86	-0.55	0.13	5	*	*
ROUSRF	0.95	-0.50	0.41	3,5	*	*
RDESRT	0.81	-0.48	0.17	3,5	*	*
MTNSNOW	0.83	-0.45	0.00	5	*	*
YTHEMT	0.59	-0.44	0.09	3,5	-.43	*
BEARTH	0.61	-0.38	0.00	3	(.86)	*
SUGARDIS	0.77	-0.35	0.00	5	*	*
PYRAMID	0.91	-0.31	0.18	5	*	*
YSOUND	0.76	-0.29	0.09	3	*	*
YINSCT	0.97	-0.29	0.17	3	*	*
YMOON	1.11	-0.17	0.30	3,5	*	*
EARTHQK	1.12	-0.14	0.23	5	*	*
YSENSE	0.81	-0.08	0.00	3	*	*

See notes at end of table.

www.manaraa.com

Table B3. Science assessment IRT item parameters: School years 2001-02 and 2003-04—Continued

Item label	IRT parameters			Used in grades	Mean and standard deviation of theta[4]	
	a[1]	b[2]	c[3]		Round 5	Round 6
THUNDER	0.87	-0.04	0.11	5	*	*
PROTECT	0.98	-0.03	0.13	5	*	*
BSHADW	0.88	-0.02	0.14	3	*	*
GRAVMOON	1.15	-0.02	0.18	5	*	*
ROUSOL	0.65	-0.01	0.16	3,5	*	*
AIRPOLL	1.17	0.07	0.00	5	*	*
YBEES	0.88	0.16	0.00	3,5	*	*
WATRGRPH	0.80	0.17	0.14	5	*	*
ROUBLB	0.69	0.23	0.20	3,5	*	*
ROUMTN	1.22	0.24	0.22	3,5	*	*
ROUGRT	0.87	0.25	0.07	3,5	*	*
ROUMCE	1.14	0.26	0.24	3,5	*	*
ROUFLY	1.03	0.28	0.13	3,5	*	*
YDSOLV	0.61	0.31	0.11	3	*	.33
LAMPWIRE	0.74	0.35	0.00	5	*	(.90)
BSOUND	0.68	0.42	0.09	3,5	*	*
MIXTURE	1.07	0.46	0.09	5	*	*
ROUSHD	0.67	0.47	0.00	3,5	*	*
YFWATE	1.09	0.47	0.24	3	*	
ECLIPSE	0.87	0.49	0.00	5	*	*
BPLNT2	0.79	0.53	0.10	3,5	*	*
YLIVE	1.16	0.62	0.12	3	*	*
BHIBER	0.56	0.65	0.18	3	*	*
CUPTEMP	0.91	0.66	0.00	5	*	*
BURIED	0.60	0.71	0.14	5	*	*
BPLANT	0.93	0.72	0.14	3,5	*	*
YBLANC	0.88	0.73	0.03	3	*	*
YFARMG	0.40	0.78	0.00	3	*	*
BSLIDE	0.88	0.98	0.10	3,5	*	*
SEEDGROW	0.74	1.04	0.04	5	*	*
H2OSOURC	0.98	1.19	0.00	5	*	*
BPLLUT	0.63	1.25	0.16	3	*	*
CHEMCHNG	0.49	1.28	0.14	5	*	*
BSOIL	1.05	1.29	0.11	3,5	*	*
BPOLAR	0.92	1.36	0.08	3		*
BSTORM	1.36	1.38	0.18	3		*
FOXRABIT	0.60	1.51	0.10	5		*
YHUMID	0.57	1.58	0.10	3		*
BPLNT3	1.03	1.60	0.06	3		*
PHYSPROP	1.05	1.83	0.24	5		*
NERVOUS	0.78	1.91	0.00	5		*
BMAMML	0.59	1.93	0.00	3,5		*
PENCLH2O	1.01	2.14	0.16	5		*

See notes at end of table.

www.manaraa.com

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

Item label	IRT parameters			Used in grades	Mean and standard deviation of theta[4]	
	a[1]	b[2]	c[3]		Round 5	Round 6
SUNMOVE	0.68	2.15	0.16	5		
CONSTELL	0.75	2.19	0.21	5		
SOLUTION	0.71	2.26	0.13	5		
TEMPLOW	0.58	2.33	0.00	5		
BEARCUB	0.34	2.58	0.18	5		
H2ORECYC	0.22	2.95	0.19	5		
WHYFAST	0.43	3.09	0.00	5		

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

www.manaraa.com

APPENDIX C
ECLS-K ESTIMATED PROPORTION CORRECT BY ROUNDS

Table C1. Reading assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, and 2003–04

	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
CANDLE	K1	0.95	0.98	0.98	0.99	1.00	1.00
POURINT	K1	0.90	0.95	0.96	0.98	1.00	1.00
CEREAL	K1	0.92	0.97	0.98	0.99	1.00	1.00
DECORATD	K1	0.86	0.92	0.94	0.97	0.99	0.99
BEGBIKE	K1	0.78	0.93	0.96	0.99	1.00	1.00
BEGIN	K1	0.67	0.82	0.86	0.94	0.98	0.99
VEGETBLE	K1	0.65	0.77	0.82	0.90	0.96	0.97
LETRECD	K1	0.72	0.93	0.96	0.99	1.00	1.00
LETRECF	K1	0.70	0.93	0.96	0.99	1.00	1.00
LETRECM	K1	0.69	0.92	0.96	0.99	1.00	1.00
LETRECT	K1	0.65	0.91	0.95	0.99	1.00	1.00
COULDNOT	K1	0.56	0.74	0.79	0.90	0.97	0.98
KAYLAFLY	K1	0.54	0.68	0.73	0.84	0.92	0.94
NEXTLINE	K1	0.54	0.75	0.81	0.92	0.98	0.99
STORYEND	K1	0.54	0.77	0.84	0.94	0.99	0.99
TIME	K1	0.53	0.73	0.79	0.91	0.97	0.98
TRUNK	K1	0.51	0.67	0.72	0.84	0.93	0.95
BEGP	K1	0.45	0.74	0.83	0.95	1.00	1.00
BEGR	K1	0.42	0.76	0.85	0.97	1.00	1.00
BEGL	K1	0.40	0.74	0.84	0.96	1.00	1.00
AWARDING	K1	0.57	0.72	0.78	0.89	0.97	0.98
JOGGING	K1	0.45	0.66	0.74	0.89	0.98	0.99
COULD	K1	0.43	0.57	0.62	0.75	0.86	0.89
ENDL	K1	0.31	0.64	0.76	0.94	1.00	1.00
MOM	K1	0.30	0.65	0.77	0.95	1.00	1.00
ENDF	K1	0.31	0.61	0.73	0.92	0.99	1.00
YELLOW	K1	0.27	0.58	0.70	0.91	0.99	1.00
BEGB	K1	0.30	0.56	0.67	0.87	0.98	0.99
BEGWORD	K1	0.35	0.53	0.61	0.78	0.91	0.94
ENDP	K1	0.26	0.53	0.65	0.88	0.98	0.99
QMARK	K1	0.30	0.52	0.61	0.81	0.95	0.97
ENDD	K1	0.21	0.46	0.59	0.85	0.98	0.99
YOU	K1	0.15	0.44	0.60	0.89	1.00	1.00

See notes at end of table.

www.manaraa.com

Table C1. Reading assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
ORPIG	K1	0.14	0.38	0.52	0.83	0.98	0.99
ORSAT	K1	0.10	0.35	0.50	0.85	0.99	1.00
ORTAIL	K1	0.07	0.29	0.44	0.83	0.99	1.00
RUNS	K1,3	0.06	0.26	0.40	0.81	0.99	1.00
ORHAND	K1	0.06	0.24	0.39	0.80	0.99	1.00
NEEDHOME	K1	0.17	0.31	0.43	0.82	0.99	1.00
WENT	K1,3	0.05	0.20	0.33	0.76	0.98	1.00
DOWN	K1,3	0.04	0.17	0.30	0.76	0.99	1.00
BOYBIRD	K1	0.22	0.33	0.43	0.79	0.99	1.00
JEEP	K1,3	0.04	0.18	0.30	0.73	0.98	0.99
GIRLRED	K1	0.09	0.24	0.34	0.65	0.91	0.96
FISHING	K1	0.03	0.13	0.25	0.74	0.99	1.00
CANINBAG	K1	0.26	0.37	0.46	0.74	0.96	0.98
KITNBED	K1	0.19	0.29	0.39	0.75	0.98	0.99
CATCH	K1	0.03	0.14	0.25	0.70	0.98	1.00
MAKE	1	0.24	0.36	0.43	0.66	0.88	0.93
KNOW	K1	0.03	0.12	0.21	0.59	0.94	0.98
LIGHT	K1	0.02	0.08	0.17	0.60	0.96	0.99
KIMCAD	K1	0.51	0.54	0.58	0.80	0.98	1.00
ELEPHANT	K1	0.02	0.08	0.16	0.58	0.96	0.99
BACKPACK	K1,3,5	0.16	0.21	0.28	0.59	0.93	0.98
LIKEDRY	K1	0.26	0.29	0.34	0.62	0.96	0.99
FLATTIRE	K1	0.16	0.20	0.26	0.56	0.94	0.98
LISTEN	K1,3,5	0.12	0.16	0.21	0.53	0.93	0.98
WRONG	K1	0.01	0.05	0.12	0.46	0.92	0.98
RIDEBIKE	K1,3,5	0.19	0.22	0.27	0.53	0.92	0.97
SIZES	K1,3,5	0.13	0.16	0.20	0.46	0.91	0.98
CHOCCAKE	K1	0.17	0.19	0.23	0.48	0.92	0.98
QUIET	K1,3	0.01	0.03	0.07	0.32	0.84	0.94
RDBIGKY	3	0.02	0.07	0.12	0.36	0.76	0.87
DOGHOUSE	K1	0.17	0.19	0.23	0.43	0.84	0.93
ENVELOPE	K1	0.01	0.03	0.07	0.30	0.84	0.94
RDFINGRY	3	0.11	0.14	0.18	0.39	0.80	0.90
RDLETR	3,5	0.28	0.29	0.32	0.48	0.85	0.94
THROUGH	K1,3,5	0.01	0.03	0.07	0.28	0.77	0.90
RDMARIAB	3,5	0.30	0.33	0.37	0.51	0.79	0.87
RDGROSR	3,5	0.20	0.22	0.26	0.43	0.79	0.89
RDLIKE	3,5	0.11	0.16	0.21	0.38	0.69	0.80

See notes at end of table.

www.manaraa.com

Table C1. Reading assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
RDDANGRY	3	0.09	0.14	0.18	0.36	0.70	0.81
RDTIME	3,5	0.33	0.34	0.36	0.48	0.84	0.93
RDENDR	3,5	0.01	0.02	0.05	0.23	0.74	0.89
RAGE	K1,3	0.00	0.02	0.05	0.22	0.75	0.90
MARCHED	K1	0.21	0.21	0.23	0.36	0.81	0.93
RDFEELSR	3,5	0.17	0.18	0.20	0.34	0.79	0.92
CATNAME	1	0.01	0.03	0.06	0.23	0.71	0.86
WTLESS	K1,3,5	0.00	0.01	0.04	0.18	0.77	0.92
RDSAMER	3,5	0.01	0.02	0.05	0.22	0.69	0.84
RDBEARY	3	0.09	0.10	0.13	0.28	0.71	0.86
TOIL	K1,3	0.01	0.03	0.05	0.22	0.68	0.84
CORNER	K1,3	0.01	0.02	0.05	0.20	0.67	0.84
RDGEORGR	3,5	0.24	0.25	0.27	0.37	0.76	0.89
OWNRNAME	1	0.01	0.02	0.05	0.20	0.66	0.83
REQUIRE	K1,3	0.00	0.01	0.03	0.15	0.69	0.87
RDTANZAR	3,5	0.00	0.01	0.04	0.17	0.67	0.85
CAPTURE	K1,3	0.00	0.02	0.04	0.17	0.65	0.83
RDFACTY	3	0.17	0.18	0.20	0.30	0.71	0.86
WEB	K1,3	0.01	0.03	0.06	0.20	0.61	0.77
RDDOCR	3,5	0.00	0.02	0.04	0.17	0.62	0.80
RDKINDY	3	0.10	0.13	0.15	0.29	0.64	0.77
UNUSUAL	K1	0.00	0.01	0.02	0.12	0.65	0.85
RDBSITY	3	0.00	0.01	0.03	0.13	0.63	0.82
MOISTURE	K1,3,5	0.00	0.01	0.03	0.13	0.63	0.82
RDSISR	3,5	0.14	0.15	0.17	0.27	0.67	0.83
MOTHER	5	0.02	0.07	0.10	0.25	0.56	0.69
RDTRUEY	3	0.11	0.12	0.14	0.25	0.63	0.79
RDSTORY	3,5	0.18	0.19	0.21	0.31	0.66	0.81
RECIPE	K1	0.19	0.20	0.21	0.28	0.67	0.84
RDSTRAGY	3	0.01	0.03	0.06	0.18	0.55	0.71
MAINPROB	5	0.15	0.18	0.20	0.32	0.60	0.73
PREDICT	5	0.13	0.13	0.15	0.22	0.62	0.80
RDWAY	3,5	0.01	0.03	0.05	0.17	0.54	0.71
RDKNIGHT	3,5	0.00	0.01	0.03	0.13	0.53	0.73
INGREDNT	K1	0.19	0.19	0.20	0.24	0.62	0.82
RDJAMEDR	3,5	0.01	0.02	0.04	0.15	0.52	0.69
EXAMPLE	5	0.14	0.14	0.16	0.23	0.59	0.77
RDCLUER	3,5	0.00	0.01	0.02	0.10	0.51	0.72

See notes at end of table.

www.manaraa.com

Table C1. Reading assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
RDBOWY	3,5	0.13	0.14	0.15	0.22	0.57	0.76
RDTRAINY	3,5	0.13	0.13	0.14	0.19	0.56	0.77
RDSUPRIR	3,5	0.00	0.01	0.02	0.10	0.49	0.70
MOREINFO	1	0.00	0.02	0.04	0.13	0.46	0.64
MYSTERLY	K1	0.00	0.00	0.01	0.06	0.45	0.69
WHYNO	1	0.02	0.04	0.07	0.19	0.47	0.61
IMP_UNDR	5	0.12	0.13	0.15	0.24	0.52	0.67
DR_ROSE	5	0.12	0.12	0.13	0.18	0.50	0.70
RDFRICTY	3	0.00	0.01	0.03	0.11	0.44	0.63
APPROX	1	0.00	0.01	0.02	0.09	0.44	0.65
RDTEARB	3,5	0.26	0.26	0.27	0.30	0.55	0.72
WAGES	K1,5	0.00	0.00	0.01	0.06	0.39	0.63
RDSAFER	3,5	0.00	0.01	0.01	0.07	0.39	0.61
MAINIDEA	1	0.29	0.30	0.31	0.36	0.58	0.71
VICIOUS	K1	0.00	0.00	0.01	0.04	0.36	0.61
RDBAKEDB	3,5	0.13	0.15	0.16	0.24	0.48	0.62
SLUDGE	5	0.10	0.10	0.11	0.16	0.42	0.61
RDPOUCHY	3	0.04	0.05	0.05	0.09	0.37	0.59
RDTHREEB	3,5	0.00	0.01	0.02	0.08	0.37	0.56
RDMOVEBY	3,5	0.06	0.08	0.10	0.19	0.43	0.56
CORNERS4	5	0.00	0.01	0.02	0.07	0.36	0.56
RDMALEBY	3	0.03	0.04	0.05	0.09	0.37	0.57
DIFFRNT	5	0.09	0.09	0.10	0.14	0.39	0.59
RDLIKER	3,5	0.00	0.00	0.01	0.05	0.33	0.54
RDDOMEST	3,5	0.00	0.00	0.01	0.05	0.33	0.54
RDAPOSTY	3	0.00	0.01	0.02	0.09	0.36	0.53
SPRING	5	0.13	0.13	0.13	0.15	0.37	0.59
RDBABONY	3	0.08	0.09	0.10	0.17	0.40	0.55
STRANDS	K1,3	0.00	0.01	0.02	0.09	0.34	0.52
SLOW_LRN	5	0.00	0.00	0.01	0.05	0.31	0.52
COMPASS	5	0.10	0.10	0.11	0.12	0.35	0.57
RDDIFFR	3,5	0.00	0.02	0.03	0.10	0.35	0.51
RDINFLUB	3,5	0.00	0.01	0.01	0.06	0.31	0.51
ABOUT	5	0.08	0.08	0.09	0.12	0.36	0.55
RDPROBLY	3,5	0.04	0.05	0.07	0.13	0.37	0.53
CRITCISM	K1,3,5	0.00	0.00	0.01	0.03	0.28	0.51
OVATIONS	5	0.24	0.25	0.26	0.30	0.49	0.62
RDBRETY	3,5	0.24	0.25	0.25	0.28	0.46	0.62

See notes at end of table.

www.manaraa.com

Table C1. Reading assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
DEPART	5	0.12	0.12	0.12	0.13	0.32	0.53
PREFRNCE	K1,3,5	0.00	0.01	0.02	0.07	0.31	0.47
WHY_LEFT	5	0.00	0.00	0.00	0.02	0.23	0.46
RDJOSHB	3,5	0.00	0.02	0.03	0.09	0.32	0.47
RDRACHLB	3,5	0.13	0.13	0.14	0.18	0.37	0.53
RDTHEMEB	3,5	0.07	0.07	0.07	0.10	0.31	0.48
WHYCONTR	5	0.06	0.06	0.07	0.10	0.28	0.45
DESCRIBE	1	0.12	0.12	0.12	0.14	0.30	0.47
RDMICROB	3,5	0.00	0.00	0.01	0.03	0.21	0.40
AMBITIO	K1,3	0.00	0.00	0.01	0.02	0.20	0.39
ON_MESA	5	0.00	0.00	0.01	0.03	0.21	0.39
RDSOLVEY	3,5	0.05	0.06	0.06	0.08	0.25	0.42
ALIGNMNT	K1,5	0.00	0.00	0.01	0.03	0.20	0.38
RDPERSNB	3,5	0.00	0.00	0.01	0.03	0.19	0.36
MTPCOMP	5	0.00	0.00	0.00	0.02	0.17	0.35
SUMMARY	5	0.10	0.11	0.12	0.18	0.35	0.46
RDHELPY	3,5	0.03	0.04	0.04	0.07	0.24	0.39
RDCOMPRB	3,5	0.00	0.00	0.01	0.04	0.20	0.36
LIKE_DIS	5	0.00	0.01	0.02	0.07	0.25	0.38
ERUPT2	5	0.00	0.00	0.01	0.02	0.17	0.33
SUPPORT	5	0.10	0.10	0.11	0.14	0.28	0.41
AUTHOR	5	0.00	0.01	0.02	0.05	0.21	0.34
PSYCHLG	5	0.00	0.01	0.02	0.05	0.20	0.32
RDGUESS	3,5	0.11	0.12	0.13	0.16	0.28	0.39
RDHOAXB	3,5	0.00	0.00	0.00	0.00	0.07	0.20
DOUBT1	5	0.00	0.00	0.00	0.00	0.04	0.17
RDCROPB	3,5	0.17	0.17	0.17	0.17	0.22	0.33
ADVANCES	5	0.01	0.02	0.03	0.07	0.22	0.32
INSUFFIC	5	0.00	0.00	0.00	0.02	0.12	0.24
DOUBT2	5	0.00	0.00	0.00	0.00	0.03	0.13
DCIRCLB	3,5	0.06	0.06	0.06	0.07	0.15	0.26
TONE	5	0.09	0.09	0.09	0.10	0.17	0.26
RDVORTXB	3,5	0.23	0.23	0.23	0.23	0.24	0.31
MAINPURP	5	0.08	0.08	0.08	0.09	0.16	0.25
THEORY2	5	0.00	0.00	0.01	0.03	0.12	0.21
RDWAGON	3,5	0.21	0.21	0.21	0.21	0.23	0.29
BELLGRNT	5	0.03	0.05	0.07	0.12	0.23	0.30
RDANOMAB	3	0.02	0.03	0.04	0.07	0.14	0.18
RDEMBOLY	3	0.00	0.00	0.01	0.01	0.04	0.07

NOTE: IRT-estimated proportion correct for each item in each round. Estimates for kindergarten through fifth grade have been put on a common scale to support comparisons. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). Not all items appeared in test forms for all rounds. Table estimates are based on cross sectional-weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0).

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

www.manaraa.com

Table C2. Mathematics assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, and 2003–04

	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
2CRAYONS	K1	0.99	1.00	1.00	1.00	1.00	1.00
3BANANAS	K1	0.89	0.95	0.96	0.98	0.99	1.00
SQUARE	K1	0.90	0.96	0.97	0.99	1.00	1.00
NUMBER 4	K1	0.90	0.98	0.99	1.00	1.00	1.00
# STRAW	K1	0.75	0.89	0.93	0.97	0.99	1.00
STICKBAT	K1	0.70	0.84	0.89	0.95	0.99	0.99
3-1PENCL	K1	0.67	0.81	0.86	0.93	0.98	0.99
NUMBER 7	K1	0.80	0.96	0.98	1.00	1.00	1.00
#VANILLA	K1	0.69	0.86	0.91	0.97	0.99	1.00
#CHOC	K1	0.62	0.82	0.89	0.96	0.99	1.00
NUMBER 9	K1	0.66	0.89	0.95	0.99	1.00	1.00
PNTBRUSH	K1	0.68	0.86	0.91	0.98	1.00	1.00
COUNT 20	K1	0.56	0.77	0.84	0.94	0.99	0.99
4LINES	K1	0.61	0.72	0.77	0.85	0.92	0.95
6BANANAS	K1	0.52	0.73	0.81	0.92	0.98	0.99
LG-SM-SM	K1	0.62	0.80	0.88	0.96	0.99	1.00
SM-LG-SM	K1	0.59	0.78	0.85	0.95	0.99	1.00
NUMBER17	K1	0.41	0.71	0.82	0.95	1.00	1.00
000X	K1	0.52	0.69	0.78	0.90	0.97	0.99
NUMBER23	K1	0.32	0.63	0.77	0.93	0.99	1.00
3RD LINE	K1	0.31	0.60	0.75	0.92	0.99	1.00
3+2 CARS	K1	0.34	0.58	0.71	0.88	0.97	0.99
_ 78910	K1	0.30	0.59	0.73	0.92	0.99	1.00
HALFOVAL	K1	0.51	0.66	0.73	0.86	0.95	0.97
2+3STICK	K1	0.30	0.55	0.69	0.88	0.98	0.99
#BUGS	K1	0.44	0.63	0.74	0.89	0.98	0.99
2 + 2	K1	0.21	0.53	0.71	0.92	1.00	1.00
3 + 3	K1	0.15	0.46	0.67	0.92	1.00	1.00
1 + 7	K1	0.24	0.47	0.61	0.82	0.96	0.98
TEAMS_R	3	0.38	0.55	0.64	0.81	0.94	0.97
VICKS_R	3	0.15	0.40	0.58	0.85	0.99	1.00
8-6CRAYN	K1	0.21	0.42	0.54	0.77	0.94	0.97
3 + 4	K1	0.12	0.34	0.51	0.81	0.98	0.99
5-1ORANG	K1	0.23	0.41	0.55	0.81	0.97	0.99
2+5MARBL	K1,3	0.17	0.35	0.48	0.73	0.93	0.97
SHAPES	K1	0.41	0.51	0.58	0.70	0.83	0.88
PATTERN	K1	0.33	0.47	0.57	0.77	0.93	0.97
2+5CIRCL	K1	0.12	0.30	0.44	0.72	0.93	0.97

See notes at end of table.

www.manaraa.com

Table C2. Mathematics assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
12 BY 2S	K1,3	0.09	0.27	0.42	0.73	0.95	0.98
3+7PENNY	K1,3	0.08	0.24	0.39	0.70	0.94	0.98
51015_25	K1,3	0.05	0.19	0.33	0.66	0.94	0.98
ORANGE_R	3	0.21	0.33	0.44	0.68	0.92	0.97
11 + 3	K1	0.05	0.18	0.31	0.64	0.93	0.98
7-3	K1	0.03	0.15	0.28	0.64	0.94	0.98
9-2	K1	0.03	0.13	0.26	0.62	0.94	0.98
PATHS_R	3	0.12	0.25	0.35	0.56	0.82	0.90
6+7	K1	0.03	0.13	0.24	0.56	0.90	0.97
12 + 6	K1	0.04	0.12	0.22	0.49	0.84	0.94
# MORE	K1	0.03	0.11	0.20	0.48	0.84	0.93
MOST_Y	3	0.02	0.08	0.17	0.47	0.87	0.96
2-1+2	K1	0.04	0.12	0.20	0.46	0.82	0.92
RULER_R	3	0.06	0.15	0.23	0.44	0.74	0.86
A13_79	K1,3,5	0.03	0.10	0.18	0.42	0.79	0.91
4+4-2	K1,3	0.02	0.07	0.14	0.40	0.81	0.93
SIDES_R	3	0.15	0.21	0.28	0.48	0.79	0.90
PAGES_R	3	0.21	0.25	0.31	0.51	0.85	0.94
17 - 4	K1	0.01	0.05	0.12	0.38	0.82	0.93
COST_10	K1,3,5	0.02	0.06	0.13	0.38	0.80	0.92
12-9	K1	0.01	0.05	0.12	0.37	0.81	0.93
26 + 20	K1	0.01	0.04	0.10	0.33	0.78	0.92
CARS15_5	K1,3,5	0.01	0.05	0.11	0.33	0.77	0.91
FEWEST_Y	3	0.01	0.04	0.09	0.30	0.75	0.90
SQUARE_R	3	0.08	0.16	0.23	0.39	0.64	0.76
CUBES10	3,5	0.08	0.16	0.23	0.38	0.63	0.76
HOWMANY$	K1,3	0.02	0.07	0.12	0.31	0.67	0.83
CANDY8_2	K1,3,5	0.01	0.03	0.08	0.26	0.70	0.87
BEADS_R	3	0.00	0.01	0.04	0.19	0.71	0.90
NEXT78	3,5	0.01	0.03	0.06	0.23	0.68	0.86
12-? PEN	K1,3	0.00	0.03	0.06	0.23	0.68	0.87
HEADSUP	K1,3	0.04	0.10	0.15	0.31	0.61	0.77
24-14BKS	K1	0.00	0.02	0.05	0.21	0.67	0.86
MEANS_R	3	0.00	0.02	0.04	0.18	0.64	0.85
EQUAL_R	3	0.17	0.18	0.20	0.31	0.71	0.88
DO_ADD4	3,5	0.01	0.03	0.06	0.21	0.63	0.83
MONEY_R	3	0.00	0.01	0.03	0.16	0.65	0.86
TIME1030	3,5	0.01	0.03	0.07	0.21	0.61	0.81

See notes at end of table.

www.manaraa.com

Table C2. Mathematics assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
POINTS_R	3	0.30	0.31	0.34	0.44	0.71	0.85
SCORE_Y	3	0.00	0.01	0.03	0.15	0.62	0.84
GOALS	K1,3	0.00	0.02	0.05	0.18	0.59	0.81
PAPERS	3	0.00	0.01	0.03	0.15	0.60	0.82
NICKELS	3	0.00	0.02	0.04	0.16	0.57	0.80
17CENTS	K1	0.00	0.01	0.03	0.13	0.56	0.80
MORE1_Y	3	0.00	0.01	0.02	0.10	0.55	0.81
NUMBER60	3,5	0.00	0.01	0.02	0.10	0.55	0.80
BDCAKE	K1	0.00	0.02	0.04	0.14	0.53	0.76
CUBESIDE	3,5	0.02	0.05	0.09	0.21	0.51	0.70
FEWER_Y	3	0.00	0.00	0.01	0.08	0.51	0.77
NEXT120	3,5	0.00	0.01	0.02	0.10	0.50	0.76
CHART_64	3,5	0.00	0.02	0.04	0.14	0.49	0.72
AGEGRAPH	5	0.01	0.03	0.06	0.17	0.49	0.70
BOX_700	3,5	0.00	0.00	0.01	0.07	0.48	0.75
NUMBER	3	0.00	0.01	0.02	0.10	0.47	0.73
SPOONS	3,5	0.00	0.01	0.02	0.09	0.47	0.73
CANDY27	5	0.00	0.01	0.01	0.08	0.46	0.73
TREES100	5	0.00	0.01	0.02	0.08	0.45	0.72
COLORSYM	3,5	0.00	0.02	0.03	0.12	0.45	0.69
FRIES	3	0.00	0.01	0.02	0.08	0.45	0.71
CHILDR_Y	3	0.00	0.01	0.02	0.08	0.44	0.71
STAR-Y	3	0.01	0.04	0.07	0.17	0.45	0.64
PAGES78	3,5	0.08	0.08	0.09	0.15	0.47	0.72
BOXSHELF	5	0.01	0.02	0.04	0.13	0.44	0.66
SECOND_Y	3	0.00	0.01	0.02	0.08	0.42	0.68
A568214K	3,5	0.00	0.01	0.01	0.07	0.41	0.69
A1ST_X5	5	0.22	0.23	0.24	0.30	0.55	0.74
PATTRN18	5	0.01	0.04	0.06	0.16	0.42	0.62
BIKETIME	5	0.00	0.01	0.02	0.08	0.39	0.65
FRUIT	3	0.12	0.13	0.14	0.20	0.46	0.68
24/4 TAB	K1	0.01	0.02	0.04	0.11	0.39	0.61
SCALE_	5	0.00	0.01	0.02	0.08	0.36	0.60
CHARGE_5	3,5	0.00	0.01	0.02	0.07	0.33	0.59
MARIA310	3,5	0.00	0.00	0.01	0.04	0.31	0.58
CARDS579	3,5	0.00	0.00	0.01	0.05	0.31	0.57
LEMONS24	5	0.00	0.00	0.01	0.04	0.29	0.55
TILES	3	0.01	0.02	0.03	0.09	0.33	0.54

See notes at end of table.

www.manaraa.com

Table C2. Mathematics assessment estimated proportion correct: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	Used in grades	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
PAIR_100	3,5	0.13	0.13	0.13	0.14	0.34	0.59
AREA_B	3	0.00	0.01	0.02	0.07	0.29	0.52
LARGER_B	3	0.00	0.01	0.01	0.05	0.27	0.50
PENCIL_Y	3	0.06	0.07	0.09	0.15	0.34	0.51
GREW4_	3,5	0.00	0.01	0.01	0.05	0.24	0.48
LOUISA13	3,5	0.00	0.00	0.00	0.02	0.19	0.46
EQUAL_B	3	0.09	0.09	0.09	0.12	0.29	0.51
AGE1_4	5	0.23	0.23	0.23	0.24	0.36	0.57
STU1_444	5	0.00	0.00	0.00	0.01	0.15	0.43
GAMESCOR	5	0.11	0.11	0.12	0.14	0.30	0.50
NUMBE2_B	3	0.00	0.00	0.01	0.02	0.17	0.40
LONGSTEP	5	0.00	0.01	0.01	0.04	0.20	0.41
MIN_BLOW	3,5	0.00	0.00	0.01	0.02	0.17	0.40
BEADSWHT	5	0.00	0.00	0.01	0.03	0.17	0.38
TALL75_	3,5	0.09	0.10	0.10	0.10	0.21	0.42
CHANGE	K1	0.00	0.00	0.01	0.03	0.16	0.37
MARBLES	3,5	0.00	0.00	0.00	0.01	0.11	0.34
BANKER_	3,5	0.00	0.00	0.01	0.03	0.16	0.35
MYSTER_B	3	0.00	0.00	0.00	0.01	0.10	0.32
OJ_30OZ	5	0.00	0.00	0.00	0.01	0.10	0.30
FRAME3FT	5	0.00	0.00	0.00	0.01	0.10	0.29
MARK_DOT	3,5	0.00	0.00	0.00	0.01	0.11	0.29
EDGECUBE	3,5	0.01	0.02	0.03	0.07	0.21	0.35
HOOP2_5	5	0.00	0.00	0.00	0.00	0.04	0.22
SAMEFRAC	3,5	0.00	0.00	0.00	0.01	0.09	0.25
SHADED_2	5	0.11	0.11	0.11	0.11	0.15	0.29
BUDGETFR	5	0.15	0.15	0.15	0.15	0.20	0.33
PIZZA	5	0.05	0.05	0.05	0.05	0.09	0.22
FRAC3_4	5	0.09	0.09	0.09	0.09	0.12	0.25
SALESTAX	5	0.20	0.20	0.21	0.23	0.31	0.41
OPOSITIV	5	0.00	0.00	0.00	0.00	0.04	0.17
AREAPLAY	5	0.05	0.05	0.05	0.05	0.10	0.21
FENCE_B	3	0.00	0.00	0.00	0.00	0.04	0.16
DIFF_88	5	0.08	0.08	0.08	0.09	0.11	0.20
SHADED_3	5	0.00	0.00	0.00	0.00	0.01	0.10
MEASDIAM	5	0.00	0.00	0.00	0.00	0.02	0.09
CARPET	5	0.00	0.00	0.00	0.00	0.00	0.04
PRISMVOL	5	0.00	0.00	0.00	0.00	0.02	0.08
TILESCOV	3,5	0.00	0.00	0.00	0.00	0.01	0.04

NOTE: IRT-estimated proportion correct for each item in each round. Estimates for kindergarten through fifth grade have been put on a common scale to support comparisons. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). Not all items appeared in test forms for all rounds. Table estimates are based on crosssectional-weights within each round (C1CW0, C2CW0, C3CW0, C4CW0, C5CW0, C6CW0).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004.

www.manaraa.com

Table C3. Science assessment estimated proportion correct: School years 2001–02 and 2003–04

	Grades	Round 5	Round 6
RBULB	3	0.90	0.95
RENRGY	3	0.94	0.98
RPLANT	3	0.94	0.98
RORGAN	3	0.79	0.86
RTOOL	3	0.84	0.92
ROUIMM	3,5	0.84	0.92
RDSAST	3	0.88	0.96
RFGRPS	3	0.78	0.86
RFORMS	3	0.82	0.91
YPLAIN	3	0.73	0.84
RWINGS	3,5	0.82	0.93
RANIML	3	0.75	0.87
ROUFRZ	3,5	0.77	0.90
ROCCUR	3	0.79	0.91
WHCHPREY	5	0.69	0.85
RSEEDS	3	0.67	0.81
ROUTAP	3,5	0.60	0.72
ROUJUN	3,5	0.67	0.84
RTHING	3	0.72	0.85
RWATER	3	0.67	0.82
YDSAST	3	0.66	0.79
RSUNIS	3	0.73	0.86
ROUERT	3,5	0.64	0.78
ROUBRN	3,5	0.62	0.81
RFISHB	3	0.63	0.79
RSHAPE	3	0.67	0.81
RHEART	3,5	0.59	0.79
RPWDER	3	0.64	0.79
ROUJAR	3,5	0.55	0.68
CUTSCAB	5	0.59	0.76
ROUSRF	3,5	0.72	0.84
RDESRT	3,5	0.60	0.76
MTNSNOW	5	0.51	0.70
YTHEMT	3,5	0.55	0.69
BEARTH	3	0.49	0.65
SUGARDIS	5	0.48	0.67
PYRAMID	5	0.56	0.74
YSOUND	3	0.51	0.69
YINSCT	3	0.55	0.73

See notes at end of table.

www.manaraa.com

Table C3. Science assessment estimated proportion correct: School years 2001–02 and 2003–04—Continued

	Grades	Round 5	Round 6
YMOON	3,5	0.59	0.76
EARTHQK	5	0.55	0.73
YSENSE	3	0.41	0.61
THUNDER	5	0.46	0.65
PROTECT	5	0.47	0.66
BSHADW	3	0.47	0.66
GRAVMOON	5	0.48	0.68
ROUSOL	3,5	0.50	0.65
AIRPOLL	5	0.34	0.59
YBEES	3,5	0.34	0.55
WATRGRPH	5	0.44	0.61
ROUBLB	3,5	0.48	0.62
ROUMTN	3,5	0.45	0.64
ROUGRT	3,5	0.36	0.56
ROUMCE	3,5	0.46	0.64
ROUFLY	3,5	0.38	0.58
YDSOLV	3	0.41	0.56
LAMPWIRE	5	0.31	0.50
BSOUND	3,5	0.37	0.53
MIXTURE	5	0.31	0.52
ROUSHD	3,5	0.30	0.47
YFWATE	3	0.42	0.60
ECLIPSE	5	0.26	0.46
BPLNT2	3,5	0.34	0.51
YLIVE	3	0.29	0.49
BHIBER	3	0.42	0.54
CUPTEMP	5	0.22	0.41
BURIED	5	0.37	0.50
BPLANT	3,5	0.31	0.48
YBLANC	3	0.23	0.41
YFARMG	3	0.32	0.43
BSLIDE	3,5	0.24	0.39
SEEDGROW	5	0.21	0.36
H2OSOURC	5	0.11	0.27
BPLLUT	3	0.30	0.41
CHEMCHNG	5	0.32	0.42
BSOIL	3,5	0.19	0.32
BPOLAR	3	0.18	0.30

See notes at end of table.

www.manaraa.com

Table C3. Science assessment estimated proportion correct: School years 2001–02 and 2003–04—Continued

	Grades	Round 5	Round 6
BSTORM	3	0.22	0.33
FOXRABIT	5	0.24	0.34
YHUMID	3	0.23	0.33
BPLNT3	3	0.12	0.22
PHYSPROP	5	0.27	0.34
NERVOUS	5	0.07	0.16
BMAMML	3,5	0.11	0.20
PENCLH2O	5	0.18	0.23
SUNMOVE	5	0.22	0.28
CONSTELL	5	0.26	0.31
SOLUTION	5	0.18	0.24
TEMPLOW	5	0.08	0.15
BEARCUB	5	0.31	0.37
H2ORECYC	5	0.37	0.41
WHYFAST	5	0.08	0.13

NOTE: IRT-estimated proportion correct for each item in each round. Estimates for third and fifth grade have been put on a common scale to support comparisons. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter). Not all items appeared in test forms for both rounds. Science was not tested in kindergarten/first grade. Table estimates are based on cross-sectional weights within each round (C5CW0, C6CW0).
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2002 and spring 2004.

www.manaraa.com

APPENDIX D

ECLS-K DIFFERENCE BETWEEN ACTUAL AND ESTIMATED PERCENT CORRECT BY ROUNDS

Table D1. Reading assessment difference between actual and estimated percent correct by rounds: School years 1998–99, 1999–2000, 2001–02, and 2003–04

	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
CANDLE	0.78	0.00	-0.01	0.00	-0.02	†	†
POURINT	0.85	0.01	-0.01	-0.03	-0.06	†	†
CEREAL	1.13	0.00	-0.01	0.00	-0.05	†	†
DECORATD	0.75	0.01	0.00	-0.01	-0.04	†	†
BEGBIKE	1.62	0.00	-0.01	0.00	-0.02	†	†
BEGIN	0.89	0.01	0.00	0.00	-0.01	†	†
VEGETBLE	0.71	0.02	-0.01	0.02	-0.08	†	†
LETRECD	2.66	0.00	0.00	-0.01	-0.01	†	†
LETRECF	3.02	0.00	0.00	-0.01	-0.01	†	†
LETRECM	2.66	0.00	0.00	-0.01	0.00	†	†
LETRECT	2.83	0.00	0.00	0.00	0.00	†	†
COULDNOT	0.88	0.01	-0.01	0.02	-0.07	†	†
KAYLAFLY	0.65	0.01	0.00	0.01	-0.06	†	†
NEXTLINE	1.10	0.00	0.00	0.03	-0.01	†	†
STORYEND	1.27	0.02	0.00	0.00	-0.02	†	†
TIME	0.99	0.02	-0.01	0.02	-0.01	†	†
TRUNK	0.71	0.02	0.01	0.00	-0.09	†	†
BEGP	1.72	0.01	0.01	0.00	-0.03	†	†
BEGR	2.30	-0.01	0.02	0.01	-0.01	†	†
BEGL	2.27	-0.01	0.02	0.02	-0.02	†	†
AWARDING	0.96	0.01	0.00	0.05	-0.02	†	†
JOGGING	1.19	0.05	0.00	0.03	-0.09	†	†
COULD	0.59	0.02	-0.01	0.01	-0.03	†	†
ENDL	2.14	0.00	0.01	0.00	-0.01	†	†
MOM	2.31	0.00	0.01	-0.01	0.01	†	†
ENDF	1.78	0.00	0.01	0.01	-0.01	†	†
YELLOW	1.88	-0.04	0.02	0.09	0.08	†	†
BEGB	1.41	0.01	0.01	0.00	-0.02	†	†
BEGWORD	0.85	0.07	0.02	0.03	-0.04	†	†
ENDP	1.61	0.01	0.00	0.01	-0.01	†	†
QMARK	1.10	-0.07	0.01	0.06	-0.01	†	†
ENDD	1.66	0.01	0.00	0.01	0.00	†	†
YOU	2.69	-0.03	0.01	0.07	0.13	†	†
ORPIG	2.07	0.06	0.02	-0.01	-0.12	†	†

See notes at end of table.

www.manaraa.com

Table D1. Reading assessment difference between actual and estimated percent correct by rounds: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
ORSAT	2.72	0.01	0.03	0.05	-0.13	†	†
ORTAIL	3.06	0.07	0.02	-0.01	-0.11	†	†
RUNS	3.33	0.00	0.00	-0.01	0.03	-0.01	†
ORHAND	3.14	0.06	0.04	0.01	-0.15	†	†
NEEDHOME	4.00	-0.01	-0.02	-0.02	0.09	†	†
WENT	3.21	0.00	0.00	-0.01	0.02	-0.01	†
DOWN	3.92	0.00	-0.02	-0.01	0.03	-0.02	†
BOYBIRD	3.63	-0.02	-0.01	0.01	0.02	†	†
JEEP	3.03	0.01	0.00	-0.01	0.02	-0.07	†
GIRLRED	1.54	0.06	-0.01	0.00	0.01	†	†
FISHING	4.86	0.00	-0.02	-0.03	0.03	†	†
CANINBAG	2.05	0.00	0.00	0.04	0.01	†	†
KITNBED	3.09	0.00	0.00	0.01	0.01	†	†
CATCH	3.67	0.01	0.00	-0.03	0.02	†	†
MAKE	1.24	†	†	0.04	-0.01	†	†
KNOW	2.53	-0.02	-0.07	-0.05	0.02	†	†
LIGHT	4.00	0.03	-0.03	-0.04	0.01	†	†
KIMCAD	4.76	0.01	0.02	0.01	0.01	†	†
ELEPHANT	3.67	0.05	0.01	0.00	-0.01	†	†
BACKPACK	2.84	0.03	0.02	0.02	-0.01	0.00	0.02
LIKEDRY	4.26	0.05	0.00	-0.03	-0.01	†	†
FLATTIRE	3.36	-0.03	-0.02	-0.04	0.02	†	†
LISTEN	3.64	0.04	0.01	0.01	0.00	0.00	-0.02
WRONG	3.50	0.05	0.05	-0.02	-0.01	†	†
RIDEBIKE	3.40	0.07	0.03	0.00	-0.01	0.00	0.00
SIZES	4.18	0.03	0.03	0.01	-0.01	0.00	-0.02
CHOCCAKE	4.93	0.05	0.00	-0.02	0.00	†	†
QUIET	3.30	†	†	0.01	0.01	-0.01	†
RDBIGKY	1.76	†	†	†	†	0.00	†
DOGHOUSE	2.64	0.02	-0.01	0.01	0.01	†	†
ENVELOPE	3.54	0.08	0.04	0.02	-0.01	†	†
RDFINGRY	2.25	†	†	†	†	0.00	†
RDLETR	2.84	†	†	†	†	0.00	0.01
THROUGH	2.56	0.06	0.00	0.01	0.02	0.01	-0.03
RDMARIAB	1.56	†	†	†	†	0.00	0.00
RDGROSR	2.13	†	†	†	†	0.01	-0.01
RDLIKE	1.33	†	†	†	†	0.00	0.04
RDDANGRY	1.49	†	†	†	†	0.00	†
RDTIME	3.23	†	†	†	†	0.00	0.02
RDENDR	2.96	†	†	†	†	-0.01	0.00
RAGE	3.33	0.06	0.02	-0.01	-0.01	0.03	†

See notes at end of table.

www.manaraa.com

Table D1. Reading assessment difference between actual and estimated percent correct by rounds: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
MARCHED	4.47	0.05	0.02	-0.02	0.00	†	†
RDFEELSR	3.72	†	†	†	†	0.00	0.00
CATNAME	2.50	†	†	0.04	0.00	†	†
WTLESS	5.21	†	†	0.00	-0.02	0.01	0.01
RDSAMER	2.45	†	†	†	†	-0.01	0.00
RDBEARY	2.42	†	†	†	†	-0.01	†
TOIL	2.38	0.06	0.02	0.01	-0.01	0.04	†
CORNER	2.48	0.03	0.00	-0.01	-0.01	0.03	†
RDGEORGR	3.22	†	†	†	†	0.02	-0.01
OWNRNAME	2.49	†	†	0.00	0.00	†	†
REQUIRE	4.22	†	†	0.02	0.00	0.00	†
RDTANZAR	3.14	†	†	†	†	0.02	-0.02
CAPTURE	2.79	0.05	0.01	-0.02	0.00	0.01	†
RDFACTY	2.98	†	†	†	†	-0.01	†
WEB	1.94	0.04	0.00	-0.01	0.00	0.01	†
RDDOCR	2.56	†	†	†	†	0.00	-0.01
RDKINDY	1.75	†	†	†	†	0.00	†
UNUSUAL	4.61	†	†	0.06	0.00	†	†
RDBSITY	3.46	†	†	†	†	-0.01	†
MOISTURE	3.44	†	†	0.00	-0.03	-0.01	0.01
RDSISR	2.86	†	†	†	†	0.00	-0.01
MOTHER	1.23	†	†	†	†	†	0.00
RDTRUEY	2.34	†	†	†	†	0.00	†
RDSTORY	2.33	†	†	†	†	0.01	-0.04
RECIPE	3.62	0.09	0.05	0.03	-0.03	†	†
RDSTRAGY	1.74	†	†	†	†	0.00	†
MAINPROB	1.46	†	†	†	†	†	0.00
PREDICT	3.07	†	†	†	†	†	-0.01
RDWAY	1.82	†	†	†	†	-0.02	0.08
RDKNIGHT	2.40	†	†	†	†	0.00	-0.03
INGREDNT	4.72	0.00	0.03	0.01	-0.02	†	†
RDJAMEDR	1.88	†	†	†	†	0.01	-0.02
EXAMPLE	2.72	†	†	†	†	†	0.00
RDCLUER	2.80	†	†	†	†	0.01	-0.01
RDBOWY	2.77	†	†	†	†	0.01	-0.01
RDTRAINY	3.57	†	†	†	†	0.01	-0.01
RDSUPRIR	2.64	†	†	†	†	-0.01	0.02
MOREINFO	1.84	†	†	-0.02	0.00	†	†
MYSTERLY	3.50	†	†	0.01	0.01	†	†
WHYNO	1.26	†	†	-0.02	0.07	†	†
IMP_UNDR	1.68	†	†	†	†	†	0.00

See notes at end of table.

www.manaraa.com

Table D1. Reading assessment difference between actual and estimated percent correct by rounds: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
DR_ROSE	2.74	†	†	†	†	†	-0.01
RDFRICTY	1.99	†	†	†	†	0.01	†
APPROX	2.44	†	†	0.00	0.00	†	†
RDTEARB	2.66	†	†	†	†	-0.11	0.02
WAGES	2.92	†	†	0.03	-0.03	†	0.00
RDSAFER	2.57	†	†	†	†	0.00	0.02
MAINIDEA	1.90	†	†	0.05	0.03	†	†
VICIOUS	3.65	†	†	0.05	0.00	†	†
RDBAKEDB	1.55	†	†	†	†	-0.02	0.01
SLUDGE	2.38	†	†	†	†	†	0.00
RDPOUCHY	2.82	†	†	†	†	0.00	†
RDTHREEB	2.02	†	†	†	†	-0.02	0.00
RDMOVEBY	1.27	†	†	†	†	0.01	-0.01
CORNERS4	2.17	†	†	†	†	†	0.00
RDMALEBY	2.30	†	†	†	†	0.00	†
DIFFRNT	2.45	†	†	†	†	†	0.00
RDLIKER	2.52	†	†	†	†	-0.01	0.04
RDDOMEST	2.43	†	†	†	†	-0.01	0.01
RDAPOSTY	1.76	†	†	†	†	0.00	†
SPRING	3.81	†	†	†	†	†	-0.01
RDBABONY	1.54	†	†	†	†	0.00	†
STRANDS	1.73	0.03	0.01	-0.01	0.00	-0.01	†
SLOW_LRN	2.42	†	†	†	†	†	0.00
COMPASS	3.50	†	†	†	†	†	0.00
RDDIFFR	1.55	†	†	†	†	0.04	-0.04
RDINFLUB	2.22	†	†	†	†	-0.03	0.01
ABOUT	2.39	†	†	†	†	†	0.00
RDPROBLY	1.55	†	†	†	†	0.04	-0.04
CRITCISM	3.16	†	†	0.01	-0.08	-0.03	0.05
OVATIONS	1.82	†	†	†	†	†	0.00
RDBRETY	2.43	†	†	†	†	0.04	-0.04
DEPART	3.63	†	†	†	†	†	-0.01
PREFRNCE	1.73	†	†	0.12	0.09	0.00	-0.02
WHY_LEFT	3.26	†	†	†	†	†	-0.01
RDJOSHB	1.49	†	†	†	†	0.08	-0.03
RDRACHLB	1.94	†	†	†	†	0.06	-0.02
RDTHEMEB	2.19	†	†	†	†	0.04	-0.02
WHYCONTR	2.07	†	†	†	†	†	0.00
DESCRIBE	2.69	†	†	-0.02	0.01	†	†
RDMICROB	2.33	†	†	†	†	-0.01	0.01
AMBITIO	2.50	†	†	0.03	0.07	0.00	†

See notes at end of table.

www.manaraa.com

Table D1. Reading assessment difference between actual and estimated percent correct by rounds: School years 1998–99, 1999–2000, 2001–02, and 2003-04—Continued

	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
ON_MESA	2.23	†	†	†	†	†	0.00
RDSOLVEY	2.21	†	†	†	†	0.02	-0.02
ALIGNMNT	2.19	†	†	-0.03	-0.11	†	0.01
RDPERSNB	2.19	†	†	†	†	0.04	-0.01
MTPCOMP	2.53	†	†	†	†	†	0.00
SUMMARY	1.26	†	†	†	†	†	0.00
RDHELPY	1.82	†	†	†	†	0.02	-0.02
RDCOMPRB	1.88	†	†	†	†	0.01	0.00
LIKE_DIS	1.36	†	†	†	†	†	0.00
ERUPT2	2.17	†	†	†	†	†	0.00
SUPPORT	1.68	†	†	†	†	†	0.00
AUTHOR	1.50	†	†	†	†	†	0.00
PSYCHLG	1.43	†	†	†	†	†	0.00
RDGUESS	1.31	†	†	†	†	0.01	0.00
RDHOAXB	3.16	†	†	†	†	-0.02	0.03
DOUBT1	4.74	†	†	†	†	†	0.00
RDCROPB	3.20	†	†	†	†	-0.02	0.03
ADVANCES	1.13	†	†	†	†	†	0.00
INSUFFIC	1.97	†	†	†	†	†	0.00
DOUBT2	4.71	†	†	†	†	†	0.00
DCIRCLB	1.91	†	†	†	†	0.02	-0.01
TONE	1.89	†	†	†	†	†	0.00
RDVORTXB	3.60	†	†	†	†	-0.01	0.02
MAINPURP	1.78	†	†	†	†	†	0.00
THEORY2	1.50	†	†	†	†	†	0.00
RDWAGON	2.49	†	†	†	†	0.04	-0.01
BELLGRNT	0.65	†	†	†	†	†	0.00
RDANOMAB	0.59	†	†	†	†	0.03	†
RDEMBOLY	0.98	†	†	†	†	0.00	†

† Not applicable.

NOTE: Difference between actual percent correct for test takers who answered each item, and IRT-estimated percent correct for the same children. Not all items appeared in test forms for all rounds. Positive numbers indicate a higher proportion of actual correct answers than were predicted by the IRT model; negative numbers correspond to actual proportions that were lower than estimates. Statistics illustrate IRT model fit, not population estimates, and are unweighted. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter).

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

www.manaraa.com

Table D2. Mathematics assessment difference between actual and estimated percent correct by rounds: School years 1998–99, 1999–2000, 2001–02, and 2003–04

	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
2CRAYONS	1.77	0.00	0.00	-0.01	-0.02	†	†
3BANANAS	0.89	0.01	-0.01	-0.01	-0.08	†	†
SQUARE	1.05	0.02	-0.02	-0.04	-0.05	†	†
NUMBER 4	3.53	-0.01	-0.01	-0.02	-0.02	†	†
# STRAW	1.21	-0.02	0.00	0.01	0.00	†	†
STICKBAT	1.01	0.03	-0.01	-0.01	-0.03	†	†
3-1PENCL	0.91	0.01	0.00	-0.02	-0.07	†	†
NUMBER 7	3.12	-0.01	-0.01	-0.03	-0.03	†	†
#VANILLA	1.35	0.00	-0.01	0.01	-0.01	†	†
#CHOC	1.43	0.00	-0.01	0.01	-0.01	†	†
NUMBER 9	2.65	-0.01	0.00	-0.01	0.00	†	†
PNTBRUSH	1.71	0.00	0.01	0.01	0.00	†	†
COUNT 20	1.28	-0.01	0.03	0.00	-0.03	†	†
4LINES	0.64	0.03	0.02	0.04	-0.07	†	†
6BANANAS	1.24	-0.01	0.00	0.03	0.01	†	†
LG-SM-SM	1.66	0.00	0.01	0.02	-0.02	†	†
SM-LG-SM	1.49	0.01	0.01	0.03	-0.01	†	†
NUMBER17	2.14	-0.02	0.02	-0.02	-0.01	†	†
000X	1.19	0.00	0.02	0.05	-0.02	†	†
NUMBER23	2.14	-0.01	0.01	-0.01	0.00	†	†
3RD LINE	2.06	0.01	0.01	0.00	-0.02	†	†
3+2 CARS	1.43	0.03	-0.01	0.00	-0.02	†	†
_ 78910	2.00	-0.03	0.02	0.04	0.01	†	†
HALFOVAL	1.01	0.00	0.02	0.02	0.00	†	†
2+3STICK	1.60	0.02	-0.01	0.01	-0.01	†	†
#BUGS	1.56	0.03	0.00	0.00	-0.02	†	†
2 + 2	3.00	-0.04	-0.01	0.02	0.03	†	†
3 + 3	4.00	-0.02	-0.01	0.00	0.01	†	†
1 + 7	1.49	0.01	-0.02	0.02	0.03	†	†
TEAMS_R	1.13	†	†	†	†	0.01	†
VICKS_R	2.42	†	†	†	†	0.00	†
8-6CRAYN	1.35	0.00	-0.01	-0.02	0.02	†	†
3 + 4	2.29	-0.05	0.00	0.01	0.00	†	†
5-1ORANG	1.99	0.03	0.00	0.00	-0.03	†	†
2+5MARBL	1.43	0.02	0.01	0.04	0.01	-0.16	†
SHAPES	0.70	0.00	0.01	0.04	-0.02	†	†
PATTERN	1.45	0.00	0.00	0.02	-0.01	†	†
2+5CIRCL	1.69	0.03	0.01	0.01	-0.03	†	†
12 BY 2S	2.08	0.01	0.00	-0.03	0.01	-0.01	†
3+7PENNY	2.07	0.02	-0.01	-0.01	0.01	-0.01	†
51015_25	2.33	-0.02	-0.01	-0.04	0.03	0.06	†
ORANGE_R	1.74	†	†	†	†	-0.01	†

See notes at end of table.

www.manaraa.com

Table D2. Mathematics assessment difference between actual and estimated percent correct by rounds: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
11 + 3	2.29	-0.01	0.00	0.01	0.01	†	†
7-3	2.82	-0.07	-0.06	-0.08	0.04	†	†
9-2	2.96	-0.07	-0.06	-0.08	0.04	†	†
PATHS_R	1.16	†	†	†	†	0.00	†
6+7	2.32	-0.01	-0.01	0.00	0.00	†	†
12 + 6	1.90	0.00	0.00	0.02	0.00	†	†
# MORE	1.97	0.10	0.05	0.06	-0.04	†	†
MOST_Y	2.68	†	†	†	†	0.00	†
2-1+2	1.80	0.04	-0.01	-0.01	0.00	†	†
RULER_R	1.29	†	†	†	†	0.00	†
A13_79	1.80	-0.02	-0.03	-0.03	0.04	0.00	-0.02
4+4-2	2.26	-0.01	0.01	0.02	0.01	-0.05	†
SIDES_R	1.61	†	†	†	†	0.00	†
PAGES_R	2.35	†	†	†	†	0.00	†
17 - 4	2.62	-0.01	-0.01	-0.02	0.02	†	†
COST_10	2.27	0.05	0.02	0.04	-0.03	0.00	-0.02
12-9	2.57	-0.03	-0.04	-0.05	0.03	†	†
26 + 20	2.67	0.00	-0.02	-0.04	0.02	†	†
CARS15_5	2.38	0.04	0.02	0.02	-0.01	-0.01	-0.02
FEWEST_Y	2.51	†	†	†	†	-0.01	†
SQUARE_R	0.96	†	†	†	†	-0.01	†
CUBES10	0.93	†	†	†	†	0.00	0.00
HOWMANY$	1.60	0.05	0.02	0.02	-0.04	0.07	†
CANDY8_2	2.34	0.02	0.02	0.04	-0.03	0.01	0.01
BEADS_R	4.06	†	†	†	†	0.01	†
NEXT78	2.50	†	†	†	†	-0.01	0.01
12-? PEN	2.55	0.06	0.04	0.03	-0.03	0.02	†
HEADSUP	1.22	0.05	0.00	0.01	-0.03	0.06	†
24-14BKS	2.77	0.02	0.02	0.02	0.00	†	†
MEANS_R	2.83	†	†	†	†	-0.01	†
EQUAL_R	2.99	†	†	†	†	0.01	†
DO_ADD4	2.23	†	†	†	†	0.02	-0.04
MONEY_R	3.44	†	†	†	†	0.01	†
TIME1030	2.02	†	†	†	†	0.00	0.01
POINTS_R	1.86	†	†	†	†	0.01	†
SCORE_Y	3.14	†	†	†	†	-0.01	†
GOALS	2.28	0.00	0.00	0.00	0.00	0.00	†
PAPERS	2.81	†	†	†	†	0.00	†
NICKELS	2.43	†	†	†	†	0.00	†
17CENTS	2.83	0.00	0.00	-0.01	0.01	†	†
MORE1_Y	3.63	†	†	†	†	0.00	†

See notes at end of table.

www.manaraa.com

Table D2. Mathematics assessment difference between actual and estimated percent correct by rounds: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
NUMBER60	3.34	†	†	†	†	0.00	-0.01
BDCAKE	2.19	0.00	0.00	-0.01	0.01	†	†
CUBESIDE	1.33	†	†	†	†	-0.01	0.01
FEWER_Y	3.62	†	†	†	†	0.00	†
NEXT120	2.83	†	†	†	†	-0.01	0.00
CHART_64	1.95	†	†	†	†	0.00	-0.01
AGEGRAPH	1.60	†	†	†	†	†	0.00
BOX_700	3.33	†	†	†	†	0.00	-0.01
NUMBER	2.48	†	†	†	†	0.00	†
SPOONS	2.72	†	†	†	†	0.00	-0.01
CANDY27	2.91	†	†	†	†	†	0.00
TREES100	2.72	†	†	†	†	†	0.00
COLORSYM	1.94	†	†	†	†	-0.06	0.03
FRIES	2.78	†	†	†	†	0.01	†
CHILDR_Y	2.60	†	†	†	†	-0.01	†
STAR-Y	1.32	†	†	†	†	0.00	†
PAGES78	2.48	†	†	†	†	0.01	-0.02
BOXSHELF	1.74	†	†	†	†	†	0.00
SECOND_Y	2.45	†	†	†	†	-0.01	†
A568214K	2.65	†	†	†	†	-0.01	0.02
A1ST_X5	1.98	†	†	†	†	†	0.01
PATTRN18	1.33	†	†	†	†	†	0.00
BIKETIME	2.16	†	†	†	†	†	-0.01
FRUIT	1.88	†	†	†	†	0.00	†
24/4 TAB	1.60	0.02	0.02	0.02	-0.01	†	†
SCALE_	1.87	†	†	†	†	†	0.00
CHARGE_5	2.05	†	†	†	†	-0.03	0.03
MARIA310	2.56	†	†	†	†	0.03	-0.02
CARDS579	2.21	†	†	†	†	-0.03	0.02
LEMONS24	2.28	†	†	†	†	†	0.00
TILES	1.49	†	†	†	†	0.00	†
PAIR_100	3.08	†	†	†	†	-0.02	0.04
AREA_B	1.70	†	†	†	†	0.00	†
LARGER_B	1.82	†	†	†	†	0.00	†
PENCIL_Y	1.18	†	†	†	†	0.00	†
GREW4_	1.85	†	†	†	†	0.02	-0.03
LOUISA13	2.67	†	†	†	†	0.03	-0.02
EQUAL_B	1.91	†	†	†	†	0.00	†
AGE1_4	2.95	†	†	†	†	†	0.00
STU1_444	3.60	†	†	†	†	†	0.00
GAMESCOR	1.91	†	†	†	†	†	0.00

See notes at end of table.

www.manaraa.com

Table D2. Mathematics assessment difference between actual and estimated percent correct by rounds: School years 1998–99, 1999–2000, 2001–02, and 2003–04—Continued

	IRT "a" parameter	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
NUMBE2_B	2.12	†	†	†	†	0.01	†
LONGSTEP	1.73	†	†	†	†	†	0.00
MIN_BLOW	2.07	†	†	†	†	0.04	-0.03
BEADSWHT	1.87	†	†	†	†	†	0.00
TALL75_	2.51	†	†	†	†	-0.05	0.02
CHANGE	1.83	0.04	0.00	0.00	0.00	†	†
MARBLES	2.69	†	†	†	†	0.00	0.03
BANKER_	1.69	†	†	†	†	0.00	-0.01
MYSTER_B	2.52	†	†	†	†	0.00	†
OJ_30OZ	2.27	†	†	†	†	†	0.00
FRAME3FT	2.25	†	†	†	†	†	0.00
MARK_DOT	1.98	†	†	†	†	0.01	-0.01
EDGECUBE	1.05	†	†	†	†	0.00	0.00
HOOP2_5	3.43	†	†	†	†	†	0.00
SAMEFRAC	1.95	†	†	†	†	-0.01	0.01
SHADED_2	2.51	†	†	†	†	†	0.00
BUDGETFR	2.15	†	†	†	†	†	0.01
PIZZA	2.52	†	†	†	†	†	0.00
FRAC3_4	2.74	†	†	†	†	†	0.00
SALESTAX	1.16	†	†	†	†	†	0.01
OPOSITIV	2.48	†	†	†	†	†	0.00
AREAPLAY	2.02	†	†	†	†	†	0.00
FENCE_B	2.17	†	†	†	†	0.00	†
DIFF_88	2.28	†	†	†	†	†	0.00
SHADED_3	3.23	†	†	†	†	†	0.01
MEASDIAM	2.10	†	†	†	†	†	0.00
CARPET	2.72	†	†	†	†	†	0.01
PRISMVOL	1.67	†	†	†	†	†	0.00
TILESCOV	1.83	†	†	†	†	0.01	0.00

† Not applicable.

NOTE: Difference between actual percent correct for test takers who answered each item, and IRT-estimated percent correct for the same children. Not all items appeared in test forms for all rounds. Positive numbers indicate a higher proportion of actual correct answers than were predicted by the IRT model; negative numbers correspond to actual proportions that were lower than estimates. Statistics illustrate IRT model fit, not population estimates, and are unweighted. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter).

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

www.manaraa.com

Table D3. Science assessment difference between actual and estimated percent correct by rounds: School years 2001–02 and 2003–04

	IRT "a" parameter	Round 5	Round 6
RBULB	0.72	0.00	†
RENRGY	1.03	0.00	†
RPLANT	1.18	-0.01	†
RORGAN	0.46	0.01	†
RTOOL	0.73	0.00	†
ROUIMM	0.81	0.00	0.00
RDSAST	1.20	0.00	†
RFGRPS	0.48	0.01	†
RFORMS	0.76	0.00	†
YPLAIN	0.62	0.00	†
RWINGS	1.31	0.02	-0.07
RANIML	0.76	0.00	†
ROUFRZ	1.17	-0.01	0.00
ROCCUR	1.26	0.01	†
WHCHPREY	0.98	†	0.00
RSEEDS	0.73	0.00	†
ROUTAP	0.48	-0.02	0.02
ROUJUN	1.12	0.00	-0.01
RTHING	0.94	0.00	†
RWATER	0.88	0.01	†
YDSAST	0.64	0.00	†
RSUNIS	0.97	0.01	†
ROUERT	0.71	0.01	-0.02
ROUBRN	1.12	0.00	-0.01
RFISHB	0.83	0.00	†
RSHAPE	0.89	0.00	†
RHEART	1.06	0.00	0.01
RPWDER	0.81	0.01	†
ROUJAR	0.49	0.00	-0.01
CUTSCAB	0.86	†	0.00
ROUSRF	0.95	0.00	0.00
RDESRT	0.81	0.01	-0.02
MTNSNOW	0.83	†	0.01
YTHEMT	0.59	0.02	-0.02
BEARTH	0.61	0.00	†
SUGARDIS	0.77	†	0.01
PYRAMID	0.91	†	0.00
YSOUND	0.76	0.00	†
YINSCT	0.97	0.00	†
YMOON	1.11	0.03	-0.03

See notes at end of table.

www.manaraa.com

Table D3. Science assessment difference between actual and estimated percent correct by rounds: School years 2001–02 and 2003–04—Continued

	IRT "a" parameter	Round 5	Round 6
EARTHQK	1.12	†	0.00
YSENSE	0.81	0.00	†
THUNDER	0.87	†	0.00
PROTECT	0.98	†	0.00
BSHADW	0.88	-0.01	†
GRAVMOON	1.15	†	0.00
ROUSOL	0.65	-0.04	0.05
AIRPOLL	1.17	†	0.00
YBEES	0.88	0.01	-0.03
WATRGRPH	0.80	†	0.00
ROUBLB	0.69	-0.01	0.02
ROUMTN	1.22	-0.02	0.03
ROUGRT	0.87	0.00	0.01
ROUMCE	1.14	0.01	-0.01
ROUFLY	1.03	0.00	0.01
YDSOLV	0.61	0.00	†
LAMPWIRE	0.74	†	0.00
BSOUND	0.68	-0.05	0.03
MIXTURE	1.07	†	0.01
ROUSHD	0.67	0.04	-0.05
YFWATE	1.09	0.00	†
ECLIPSE	0.87	†	0.00
BPLNT2	0.79	-0.02	0.01
YLIVE	1.16	0.01	†
BHIBER	0.56	0.00	†
CUPTEMP	0.91	†	0.01
BURIED	0.60	†	0.00
BPLANT	0.93	0.08	-0.02
YBLANC	0.88	0.00	†
YFARMG	0.40	0.00	†
BSLIDE	0.88	-0.06	0.02
SEEDGROW	0.74	†	0.00
H2OSOURC	0.98	†	0.00
BPLLUT	0.63	0.00	†
CHEMCHNG	0.49	†	0.00
BSOIL	1.05	-0.03	0.02
BPOLAR	0.92	0.01	†
BSTORM	1.36	0.01	†
FOXRABIT	0.60	†	0.00
YHUMID	0.57	0.00	†

See notes at end of table.

www.manaraa.com

Table D3. Science assessment difference between actual and estimated percent correct by rounds: School years 2001–02 and 2003–04—Continued

	IRT "a" parameter	Round 5	Round 6
BPLNT3	1.03	0.01	†
PHYSPROP	1.05	†	0.01
NERVOUS	0.78	†	0.01
BMAMML	0.59	0.05	-0.02
PENCLH2O	1.01	†	0.01
SUNMOVE	0.68	†	0.01
CONSTELL	0.75	†	0.01
SOLUTION	0.71	†	0.01
TEMPLOW	0.58	†	0.00
BEARCUB	0.34	†	0.00
H2ORECYC	0.22	†	0.00
WHYFAST	0.43	†	0.00

† Not applicable.

NOTE: Difference between actual percent correct for test takers who answered each item and IRT-estimated percent correct for the same children. Not all items appeared in test forms for all rounds. Positive numbers indicate a higher proportion of actual correct answers than were predicted by the IRT model; negative numbers correspond to actual proportions that were lower than estimates. Statistics illustrate IRT model fit, not population estimates, and are unweighted. Items are sorted in estimated ascending order of overall difficulty (IRT "b" parameter).

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

www.manaraa.com