# The Illinois State Assessment
# 1999 Technical Manual

**Illinois State Board of Education**
**Standards and Assessment Division**

## CONTENTS

# Part 1: The Illinois Standards Achievement Tests

# 1. PURPOSE AND DESIGN OF THE ISAT TESTING PROGRAM

In February 1999, students in grades 3, 5, 8, and 10 took Illinois Standards Achievement Tests (ISAT) in reading, mathematics, and writing. More than 500,000 students enrolled in public elementary and secondary schools across the state participated in the testing program. ISAT measures the extent to which students are meeting the Illinois Learning Standards (1997). Illinois teachers and curriculum experts developed the ISAT tests in cooperation with the Illinois State Board of Education (ISBE).

This manual provides technical information about the 1999 test administration. It describes the tests and assessment approaches and addresses technical concerns. Other reports, documents, or publications issued by the Illinois State Board of Education (ISBE) provide additional information about interpreting test results (*Guide to the 1999 Illinois State Assessment*, *Understanding Your Child's ISAT Scores*) that is not included here.

## General Procedures

Each ISAT test is designed to ensure that its results validly and fairly assess the Illinois Learning Standards. The selection of items and assembly of each test is guided by a set of specifications. These specifications were developed by Illinois educators to help ensure that test content corresponds to the purposes, objectives, and skills framed by the learning standards.

Illinois teachers and administrators participate in all phases of the test development process: item writing, item selection, bias review, and test assembly. The State Board of Education convenes a series of advisory committees to ensure that test development is continually informed and guided by the recommendations of content authorities, measurement specialists, and practitioners. The following evaluation criteria are applied to all assessment material used in the Illinois program:

*Content.* Every item is screened for alignment with the Illinois Learning Standards, grade-level appropriateness, importance, and clarity. Incorrect choices (for multiple-choice items) are reviewed for plausibility. In tests other than reading, the complexity of the text of the questions is kept to the minimum necessary to state the problem.

*Difficulty.* Items are pilot tested on large samples of students prior to their inclusion in tests to develop a statistical profile for each item. Items that are too easy or too difficult and, therefore, provide little or no information are omitted.

*Precision.* Point-biserial (i.e., item-test) correlations evaluate the extent to which an item distinguishes between less proficient and more proficient students. Reviewers usually omit items with a point-biserial of less than .30 and select items with the highest point-biserial.

*Fairness.* Test items and forms undergo regular sensitivity reviews and statistical analyses to ensure that all materials meet fairness criteria with respect to the cultural and ethnic diversity of Illinois public schools.

ISBE takes several precautions to help ensure test security. Test materials shipped to schools are packaged and sealed. Each test booklet is barcoded so that it can be accounted for. The administration of tests is standardized. A series of manuals provides guidance on security and other issues to the district testing coordinator, school testing coordinator, and classroom test administrator. After administration, all materials are removed from schools and returned to a central facility for processing and secure destruction of unneeded materials.

## Reading

The ISAT reading test assesses material defined by standards associated with three state learning goals. The standards were developed using the 1985 State Goals for Language Arts, various state and national standards drafts, and local education standards contributed by team members. These learning standards are designed to guide language arts instruction in Illinois schools. This alignment of assessment to curriculum insures consistency and strengthens the influence of standards and assessment on improved teaching and learning. These standards are:

- Goal 1: Read with understanding and fluency.
  - 1A. Apply word analysis and vocabulary skills to comprehend selections.
  - 1B. Apply reading strategies to improve understanding and fluency.
  - 1C. Comprehend a broad range of reading materials.

- Goal 2: Read and understand literature representative of various societies, eras and ideas.
  - 2A. Understand how literary elements and techniques are used to convey meaning.
  - 2B. Read and interpret a variety of literary works.

- Goal 5: Write to communicate for a variety of purposes.
  - 5A. Locate, organize, and use information from various sources to answer questions, solve problems and communicate ideas.
  - 5B. Analyze and evaluate information acquired from various sources.
  - 5C. Apply acquired information, concepts and ideas to communicate in a variety of formats.

The reading test has two formats. The grade 3 reading assessment is given in three 35-minute sessions. One of these sessions consists of 12-15 word analysis questions and one passage followed by 15-17 multiple-choice questions. The two remaining sessions include one passage followed by 15-20 multiple-choice questions, and one short answer question.

The reading tests for grades 5, 8, and 10 are also given in three 35-minute sessions. One of these sessions consists of a longer passage with 15-20 multiple-choice questions. The other

two sessions each include one passage with 15-20 multiple choice questions and one short answer question.

The reading passages and accompanying questions reflect two of the most frequent purposes for reading—reading to gain information and reading for literary experience. The sources for these passages range from high interest, grade-appropriate periodicals to newspapers, short stories, and novels. Illinois teachers reviewed and selected the material for these tests.

The multiple-choice questions require students to select one correct response from four possibilities presented to them. Again, teachers in Illinois played an active part in writing, reading, and editing these test questions. Questions must meet both content and statistical criteria for inclusion in the test.

The short answer questions on the reading test require students not only to read and understand a text, but also to analyze, evaluate, and interpret the text as a means of making connections and conclusions related to the text. The rubric used to score the short answer responses is a holistic scoring rubric. It describes characteristics of different levels of achievement in reading. The levels of achievement on the reading rubric range from 0 to 4 (4 being the highest score). Responses with scores of 0 indicate that the student response is insufficient to effectively determine evidence of achievement in reading. Responses with scores of 1 and 2 indicate developing levels of achievement in reading. Responses with scores of 3 indicate a developed level of achievement in reading. Finally, responses with scores of 4 represent a well-developed level of achievement in reading. The rubric was developed with Illinois educators. The reporting of the short answer item scores is different than that of the other questions. For the first two years of the assessment, schools and districts will receive the short answer item scores for informational purposes rather than accountability purposes.

In addition to an overall reading score, results are reported in terms of the percent of items correctly answered within five "standard sets" (six at grade 3). These scores are as follows:

- *Comprehension: Literary Works:* Understanding of passages taken from sources such as novels, short stories, and periodicals. (Standards 1B, 1C, 2A, 2B, 5A, 5B, 5C)

- *Comprehension: Informational Sources:* Understanding of non-fiction texts such as student periodicals, newspapers, and trade journals. (Standards 1B, 1C, 2A, 2B, 5A, 5B, 5C)

- *Application of Strategies: Explicit Ideas:* Identifying important information directly stated in the text. (Standards 1B, 5A)

- *Application of Strategies: Inferences from Text:* Analyzing important information in the text to draw logical conclusions about the text. (Standards 1C, 2A, 2B, 5B, 5C)

- *Vocabulary:* Using contextual clues and other skills to understand key words, phrases, and concepts in literary and informational texts. (Standard 1A)

- *Word Analysis (3rd grade only):* Using phonics, word pattern, and other word analysis skills to recognize new words. (Standard 1A)

## Mathematics

People use mathematics to identify, describe and investigate the patterns and challenges of everyday living. Mathematics helps us to understand events that have occurred and to predict and prepare for events to come so that we can more fully understand our world and more successfully live in it. Mathematics encompasses arithmetic, measurement, algebra, geometry, trigonometry, statistics, probability and other fields. It deals with numbers, quantities, shapes and data, as well as numerical relationships and operations. Confronting, understanding and solving problems is at the heart of mathematics. Mathematics is much more than a collection of concepts and skills; it is a way of approaching new challenges through investigating, reasoning, visualizing and problem-solving with the goal of communicating the relationships observed and problems solved to others.

The ISAT mathematics tests are designed to measure the following learning standards.

- Goal 6: Demonstrate and apply a knowledge and sense of numbers, including numeration and operations (addition, subtraction, multiplication, division), patterns, ratios and proportions.
  - 6A. Demonstrate knowledge and use of numbers and their representations in a broad range of theoretical and practical settings.
  - 6B. Investigate, represent and solve problems using number facts, operations (addition, subtraction, multiplication, division) and their properties, algorithms and relationships.
  - 6C. Compute and estimate using mental mathematics, paper-and-pencil methods, calculators and computers.
  - 6D. Solve problems using comparison of quantities, ratios, proportions and percents.

- Goal 7: Estimate, make and use measurements of objects, quantities and relationships and determine acceptable levels of accuracy.
  - 7A. Measure and compare quantities using appropriate units, instruments and methods.
  - 7B. Estimate measurements and determine acceptable levels of accuracy.
  - 7C. Select and use appropriate technology, instruments and formulas to solve problems, interpret results and communicate findings.

- Goal 8: Use algebraic and analytical methods to identify and describe patterns and relationships in data, solve problems and predict results.
  - 8A. Describe numerical relationships using variables and patterns.
  - 8B. Interpret and describe numerical relationships using tables, graphs and symbols.
  - 8C. Solve problems using systems of numbers and their properties.
  - 8D. Use algebraic concepts and procedures to represent and solve problems.

- Goal 9: Use geometric methods to analyze, categorize and draw conclusions about points, lines, planes and space.

9A. Demonstrate and apply geometric concepts involving points, lines, planes and space.
9B. Identify, describe, classify and compare relationships using points, lines, planes and solids.
9C. Construct convincing arguments and proofs to solve problems.
9D. Use trigonometric ratios and circular functions to solve problems.

- Goal 10: Collect, organize and analyze data using statistical methods; predict results; and interpret uncertainty using concepts of probability.
  10A. Organize, describe and make predictions from existing data.
  10B. Formulate questions, design data collection methods, gather and analyze data and communicate findings.
  10C. Determine, describe and apply the probabilities of events.

Illinois teachers developed the Illinois Learning Standards for Mathematics. These goals, standards and benchmarks are an outgrowth of the 1985 Illinois State Goals for Learning influenced by the latest thinking in school mathematics. This includes the National Council of Teachers of Mathematics; Curriculum and Evaluation Standards for School Mathematics; ideas underlying recent local and national curriculum projects; results of state, national, and international assessment findings; and the work and experiences of Illinois school districts and teachers.

The mathematics assessment includes 80 multiple-choice items at grade 3 and 90 multiple-choice items at the other grades administered in three test sessions. In addition, the tests contain two short-answer/problem-solving tasks. As is true in reading, for the first two years of the assessment, schools and districts will receive the short answer item scores for informational purposes rather than accountability purposes.

In addition to an overall mathematics score, results are reported in terms of the percent of items correctly answered within eight standard sets. Not all sets are reported at all grades. These scores are as follows:

- *Estimation/Number Sense/Computation:* Demonstrating an understanding of numbers, their representations, and number operations of addition, subtraction, multiplication, division, percentages, and fractions as appropriate to grade level. (Standards 6A, 6B, 6C, 6D, 8C)

- *Algebraic Patterns/Variables:* Identifying, describing, and extending algebraic, geometric, and numeric patterns and constructing and solving problems using variables. (Standards 8A, 8C, 8D)

- *Algebraic Relationships/Representations:* Representing and interpreting algebraic concepts with words, diagrams, tables, coordinate graphs, equations, and inequalities. (Standards 8B, 8C)

- *Geometric Concepts:* Identifying and describing points, lines, two- and three-dimensional shapes and their properties, such as parallel; symmetry; perpendicular; and number of sides, faces, and vertices. (Standards 8C, 9A)

- *Geometric Relationships:* Sorting, classifying, comparing, and contrasting geometric figures. This category includes such properties as similarity and congruency. (Standards 8C, 9B, 9D)

- *Measurement:* Estimating, measuring, and comparing quantities using appropriate units and acceptable levels of accuracy. At higher grades, this category encompasses conversions within measurement systems. (Standards 7A, 7B, 7C, 8C)

- *Data Organization/Analysis:* Creating, analyzing, displaying, and interpreting data using a variety of graphs (pictures, tallies, tables, charts, bar graphs, Venn diagrams), and computing the mean, median, mode, and range of given data. (Standards 8C, 10A, 10B)

- *Probability:* Determining, describing, and applying elementary probability theory and fundamental counting principles. At higher grades, this category encompasses combinations and permutations of simple and complex events. (Standard 10C)

## Writing

The state goal for writing states that the student will be able to write standard English in a grammatical, well-organized, and coherent manner for a variety of purposes. The learning standards associated with the goal are as follows:

3A. Use correct grammar, spelling, and punctuation
3B. Compose well-organized and coherent writing
3C. Communicate ideas in writing to accomplish a variety of purposes

The writing assessment uses three types of prompts, which represent persuasive, expository, and narrative discourse modes. Persuasive topics require students to take a position on an issue or to state a problem and solution. Expository topics require students to explain, interpret, or describe something objectively and clearly. Narrative topics require students to reflect upon and describe an experience or event from personal knowledge. Readers evaluate each paper with respect to its focus, support/elaboration, organization, and conventions. They also evaluate how effectively the paper integrates these features.

Students in grades 5, 8, and 10 wrote one assigned essay. All students within a grade received the same assignment. They then selected a second topic (or prompt) from a list of two and wrote a second essay. Third-grade students received one of three topics and wrote an essay on the assigned topic.

Readers score all papers with respect to four specific features (focus, support/elaboration, organization, and conventions) and a holistic feature (integration). Descriptions of these features follow:

- *Focus:* the degree to which the subject, issue, theme, or unifying event of the composition is clear and maintained.

- *Support/Elaboration:* the quality of the detail or support through reasons and explanations.

- *Organization:* the extent to which a clear structure or plan of development is maintained and the points logically related to each other and the text structure.

- *Conventions:* the extent to which the writer demonstrates adequate knowledge of standard English.

- *Integration:* the extent to which the paper as a whole uses the four features (focus, support, organization, and conventions) to address the assignment.

Readers rate a paper's first three features and its overall integration on a scale from 1 (absent) to 6 (well developed). The conventions feature is evaluated as either 1 (not developed) or 2 (developed). A composite writing score is derived from the raw feature scores according to the following formula:

Focus + Support + Organization + Conventions + (2 x Integration)

The overall writing score ranges from 6 to 32. For students who wrote more than one essay (grades 6, 8, 10), writing scores for each essay were averaged and then rounded up. Thus, individual student scores at all grades are reported as whole numbers. Scores for schools, districts, and the state are reported to one decimal place.

# 2. RELIABILITY

The reliability of a test reflects the degree to which scores are free from random errors of measurement. Test reliability indicates the extent to which differences in test scores reflect real differences in the ability being measured and, thus, the consistency of test scores across some change of condition, such as a change of test items or a change of time. Different reliability coefficients result from different changes in testing conditions. For example, test-retest reliability measures the extent to which scores remain constant over time. A low test-retest reliability coefficient means that a person's scores are likely to shift unpredictably from one time to another.

## Internal Consistency of Overall Scores

Because the items used in achievement tests represent only a relatively small sample from a much larger domain of items, the consistency of test scores across items is of particular interest. That is, how precisely will tests rank students if different sets of items from the same domain were used? Unless the rankings are very similar, it is difficult or impossible to make educationally sound decisions on the basis of test scores. This characteristic of test scores is most commonly referred to as *internal consistency*. Table 2.1 presents internal consistency values (coefficient alpha) for each of the tests administered in the 1999 assessment.

**Table 2.1**
**1999 Reliability Estimates**

| Grade | Reading | Mathematics | Writing |
|---|---|---|---|
| 03 | .95 | .95 | .87 |
| 05 | .93 | .95 | .90 |
| 08 | .93 | .95 | .90 |
| 10 | .95 | .96 | .91 |

Note: Sample sizes on which these coefficients are based are as follows:

Reading, Mathematics: 3 (145,105), 5 (139,753), 8 (137,178), 10 (120,150)
Writing: 3 (140,003), 5 (135,339), 8 (133,169), 10 (113,873)

The reliability coefficients reported in Table 2.1 are derived within the context of classical test theory (CTT) and provide a single measure of precision for the entire test. Within the context of item response theory (IRT), it is possible to measure the relative precision of the test at different points on the scale. Figure 2.1 presents the test information functions for the four ISAT reading tests; Figure 2.2 presents comparable information for the four ISAT mathematics tests. IRT scaling is not used with the writing test.

The amount of information at any point is directly related to the precision of the test. That is, precision is highest where information is highest. Conversely, where information is lowest, precision is lowest and ability is most poorly estimated.

**Figure 2.1**
**1999 ISAT Reading Test Information Functions**

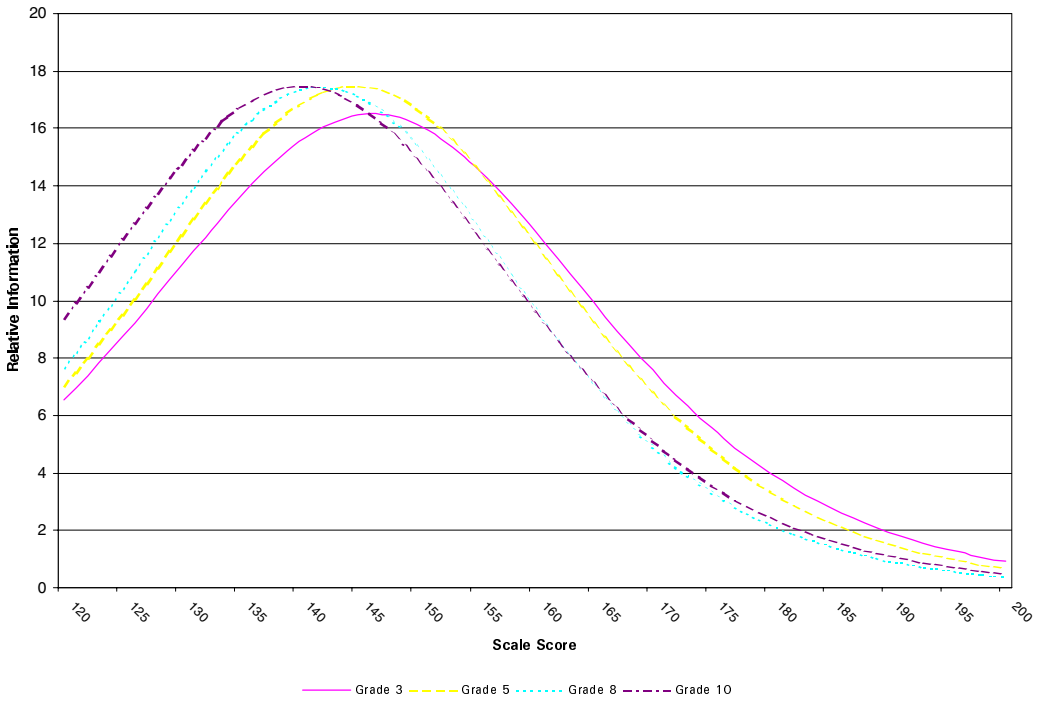
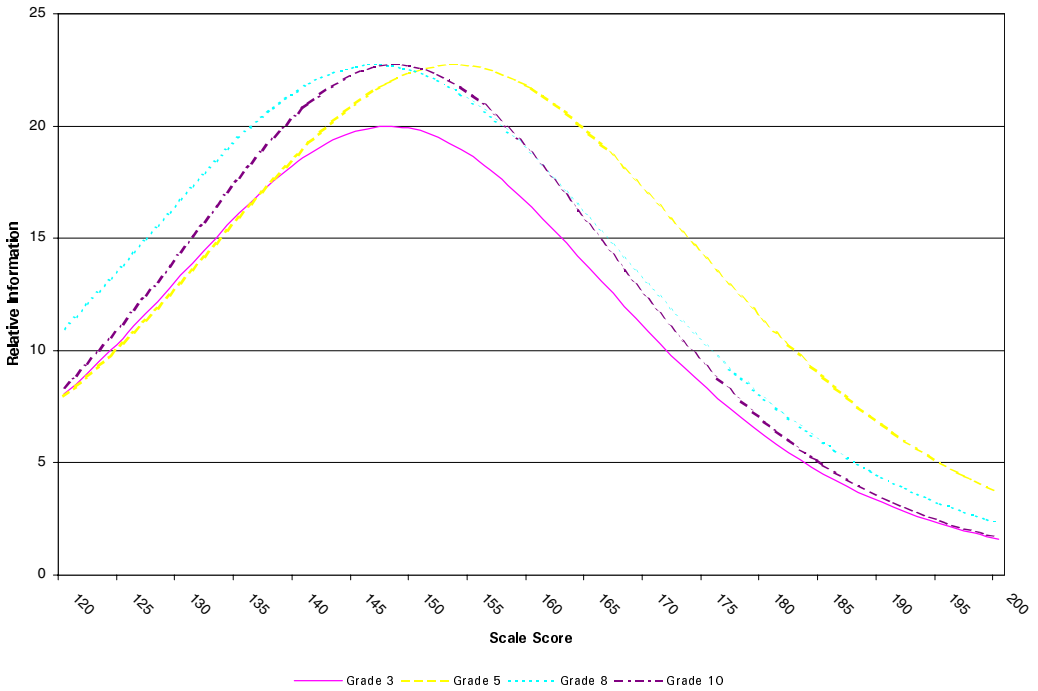



**Figure 2.2**
**1999 ISAT Mathematics Test Information Functions**



A second way of evaluating precision from the IRT perspective is in terms of how well the test as a whole separates people. The ratio of the standard deviation of ability estimates after subtracting from their observed variance the error variance attributable to their standard errors of measurement to the root mean square standard error computed over persons provides this index (Wright & Stone, 1979). These values are reported in Table 2.2.

**Table 2.2**
**Person Separation Values for the ISAT Reading and Mathematics Tests**

| | |
|---|---|
| Reading | |
| Grade 3 | 3.14 |
| Grade 5 | 2.71 |
| Grade 8 | 2.42 |
| Grade 10 | 2.60 |
| | |
| Mathematics | |
| Grade 3 | 3.61 |
| Grade 5 | 3.73 |
| Grade 8 | 3.85 |
| Grade 10 | 4.06 |

## Reliability of the Writing Scores

Writing scores are affected by other sources of variance, particularly readers (raters), since different readers evaluate different students, and prompts. The effect attributable to prompts is important for students at all grades. However, it can only be evaluated directly for 5th-, 8th- and 10th-grade students who wrote on two different prompts.

*Interrater Agreement.* Interrater agreement evaluates the consistency of scores assigned to the same essay by different readers. For the 1999 writing assessment, interrater agreement was monitored daily, and two readers independently scored 10% of the student essays across grades and prompts. The interrater agreement coefficients for all features and discourse modes are summarized in Table 2.3. The results for the interrater agreement on double-scored papers exceeded the minimum acceptable level of agreement (90% agreement within one point). Scores across raters agree within one point at least 93% of the time.

**Table 2.3**
**Interrater Agreement for Writing Scores**

| Discourse Mode | Score | % Exact Agreement | % Adjacent Agreement | % Exact + Adjacent |
|---|---|---|---|---|
| Persuasive | Focus | 60 | 35 | 95 |
| | Support | 52 | 42 | 94 |
| (n = 38,815) | Organization | 53 | 42 | 95 |
| | Conventions | 93 | 7 | 100 |
| | Integration | 56 | 40 | 96 |
| Expository | Focus | 62 | 34 | 96 |
| | Support | 53 | 42 | 95 |
| (n = 30,372) | Organization | 54 | 41 | 95 |
| | Conventions | 93 | 7 | 100 |
| | Integration | 58 | 39 | 97 |
| Narrative | Focus | 54 | 39 | 93 |
| | Support | 54 | 40 | 94 |
| (n = 8,174) | Organization | 54 | 40 | 94 |
| | Conventions | 92 | 8 | 100 |
| | Integration | 56 | 39 | 95 |

In addition to agreement across raters, writing scores are checked against a standard or "validation" set of papers. The Validation Committee assigns the scores for these papers. Essay packets, each containing 10 essays, were circulated among the readers. Essays for these check sets were chosen to represent a range of score points in all categories.

Readers encountered the validation packets at random intervals throughout the scoring, and some encountered several packets during the scoring process. Readers were unaware of the scores assigned to the papers by the committee. The extent of agreement between a reader's scores and the scores assigned to the papers was calculated every day during the scoring and shared with the readers. This process allowed for the monitoring of reader scoring. The results for all features and discourse modes are summarized in Table 2.4. Again, the results exceeded the minimum acceptable level of agreement (90% agreement within one point). The agreement of readers with validation papers was higher than the interrater agreement. This is possibly attributable to the fact that the validation papers are specifically selected to illustrate all points on the scoring scale. The papers that are selected for double scoring, on the other hand, represent a more nearly random selection of papers and scores. Consequently, they are likely to include proportionately fewer extreme scores (e.g., 1, 6), on which there is likely to be higher agreement between raters.

**Table 2.4**
**Agreement with Validation Papers for Writing Scores**

| Discourse Mode | Score | % Exact Agreement | % Adjacent Agreement | % Exact + Adjacent |
|---|---|---|---|---|
| Persuasive/ Expository | Focus | 69 | 29 | 98 |
| | Support | 65 | 33 | 98 |
| | Organization | 65 | 33 | 98 |
| (N = 7,520) | Conventions | 91 | 9 | 100 |
| | Integration | 67 | 32 | 99 |
| Narrative | Focus | 62 | 34 | 97 |
| | Support | 57 | 39 | 97 |
| (N = 2,380) | Organization | 62 | 34 | 96 |
| | Conventions | 81 | 19 | 100 |
| | Integration | 64 | 34 | 97 |

## Reliability of the Performance Category Decisions

Students' ISAT scores are reported relative to four performance categories: Academic Warning, Below Standards, Meets Standards, and Exceeds Standards. Sets of score cutoffs that were developed for each learning area and each grade. The development of the score cutoffs that define these categories is fully documented in a separate publication available from ISBE (*Performance Levels for the Illinois Standards Achievement Tests*). However, the process may be briefly described as follows.

Prior to the meetings of the standard-setting panels themselves, which took place during April 1999, ISBE convened committees of curriculum experts to develop concrete descriptions of student knowledge and skill levels that define the specific performance categories. Educators throughout Illinois extensively reviewed these descriptions.

Panels of recognized subject matter experts convened in Springfield to translate the verbal descriptions into cut scores on the ISAT tests (i.e., scores that define the boundaries between categories). Panelists were drawn from a pool of educators who had specific knowledge of student performance at the grade levels being assessed by ISAT and experience in assessing students at those grade levels. Panelists were selected to be broadly representative of the geographic and ethnic diversity of Illinois' public school system. A total of 170 educators participated in the standard-setting process. The distribution of educators across learning areas was as follows: mathematics—56; writing—62; reading—52.

A procedure originally proposed by Angoff is one of the most frequently used methods for determining cut scores when multiple-choice test scores are used. It can be most simply described as a focused, judgmental process by knowledgeable content experts. The basic Angoff procedure fit the format of the ISAT reading and mathematics tests. However,

certain modifications of the basic procedure were developed to fit the format of the ISAT writing tests.

In the most frequent application of the Angoff method (e.g., to establish a pass-fail standard), panelists are asked to examine an item and decide what proportion of minimally competent individuals will answer the question correctly. With respect to the ISAT, however, instead of being asked about minimally competent students, panelists were asked to indicate what percentage of three groups of students—those who were just above the Academic Warning/Below Standards boundary, those who were just above the Below Standards/Meets Standards boundary, and those who were just above the Meets Standards/Exceeds Standards boundary—would answer the question correctly. The ratings were made sequentially rather than simultaneously (i.e., panelists made all judgments relative to one cut score before moving to the next cut score). Item performance statistics were provided to help panelists anchor their ratings. The cutoff scores that resulted are shown in Table 2.5. Results of applying these cutoffs to the 1999 test population are shown later in Section 5.

**Table 2.5**
**ISAT Cutoffs for Each Performance Level**

| READING | | | | |
|---|---|---|---|---|
| | Academic Warning | Below Standards | Meets Standards | Exceeds Standards |
| 03 | 120-137 | 138-155 | 156-173 | 174-200 |
| 05 | 120-129 | 130-155 | 156-170 | 171-200 |
| 08 | 120-128 | 129-151 | 152-172 | 173-200 |
| 10 | 120-135 | 136-152 | 153-174 | 175-200 |
| MATHEMATICS | | | | |
| | Academic Warning | Below Standards | Meets Standards | Exceeds Standards |
| 03 | 120-141 | 142-152 | 153-172 | 173-200 |
| 05 | 120-137 | 138-157 | 158-190 | 191-200 |
| 08 | 120-137 | 138-161 | 162-184 | 185-200 |
| 10 | 120-138 | 139-157 | 158-187 | 188-200 |
| WRITING | | | | |
| | Academic Warning | Below Standards | Meets Standards | Exceeds Standards |
| 03 | 6-13 | 14-21 | 22-29 | 30-32 |
| 05 | 6-13 | 14-20 | 21-27 | 28-32 |
| 08 | 6-14 | 15-20 | 21-27 | 28-32 |
| 10 | 6-14 | 15-20 | 21-27 | 28-32 |

The reliabilities of such classifications, which are criterion-referenced, are related to the reliabilities of the tests on which they are based, but they are not equivalent to the test

reliabilities, which are based on norm-referenced measurement. Glaser (1963) was among the first to draw attention to this distinction, and Feldt and Brennan (1989) extensively reviewed the topic.

As Feldt and Brennan (1989, p. 140) point out, approaches to the development of reliability coefficients for criterion-referenced interpretations of test scores have been based either on squared-error loss or threshold loss. It is threshold loss, which evaluates the consistency with which people are consistently classified with respect to a criterion, that is of greater concern here. Specifically, the issue is how consistently do tests classify students with respect to the performance standards?

Two threshold-loss coefficients have been developed: $p$, the proportion of persons consistently classified on two parallel tests, and $k$ (kappa), which corrects $p$ for the proportion of consistent classifications that would be expected by chance. Because scores on classically parallel tests are rarely available in practice, methods have been developed to estimate these values from a single test (Subkoviak, 1984). An approach proposed by Peng and Subkoviak (1980) was applied to the performance classifications made on the basis of the 1999 tests.

Table 2.6 presents the 1999 values for $p$, $k$, and $p_{miss}$, the expected proportion of inconsistent decisions, which is simply (1 - $p$). In interpreting the first two indexes, Feldt and Brennan (1989) suggest that $p$ reflects the *consistency of decisions* made about examinees, whereas $k$, since it is corrected for chance, reflects the *contribution of the test* to the consistency of the decision.

Overall, the values suggest that decisions made with respect to the student performance classifications would be very consistent. Note that the $p$ and $k$ values are calculated for the complete test population. Values for other test populations (e.g., IEP students alone, non-IEP students only) may differ.

**Table 2.6**
**Reliability of Student Performance Decisions Based on 1999 Test Scores**

| | | Academic Warning/Below Standards | | | Below Standards/Meets Standards | | | Meets Standards/Exceeds Standards | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Area | Grade | p | kappa | $p_{miss}$ | p | kappa | $p_{miss}$ | p | kappa | $p_{miss}$ |
| Reading | 3 | .970 | .738 | .030 | .906 | .797 | .094 | .930 | .780 | .070 |
| | 5 | .990 | .682 | .010 | .866 | .710 | .134 | .916 | .789 | .084 |
| | 8 | .994 | .724 | .006 | .880 | .698 | .120 | .896 | .689 | .104 |
| | 10 | .964 | .614 | .036 | .878 | .704 | .122 | .910 | .683 | .090 |
| Mathematics | 3 | .954 | .757 | .046 | .974 | .933 | .026 | .928 | .785 | .072 |
| | 5 | .970 | .738 | .030 | .924 | .783 | .076 | .930 | .780 | .070 |
| | 8 | .972 | .731 | .028 | .924 | .783 | .076 | .942 | .769 | .058 |
| | 10 | .964 | .737 | .036 | .886 | .713 | .114 | .982 | .709 | .018 |
| Writing | 3 | .952 | .538 | .048 | .824 | .646 | .176 | .936 | .570 | .064 |
| | 5 | .998 | .799 | .002 | .896 | .689 | .104 | .888 | .695 | .112 |
| | 8 | .976 | .574 | .024 | .864 | .712 | .136 | .958 | .633 | .042 |
| | 10 | .970 | .609 | .030 | .872 | .700 | .128 | .914 | .678 | .086 |
| AVERAGE | | .973 | .687 | .027 | .891 | .739 | .109 | .928 | .713 | .073 |

# 3. ISAT SCALING PROCEDURES

Because test items change each year, raw scores (i.e., number or percent correct scores) will not always have the same meaning or represent the same level of proficiency across forms. This is attributable to variations in difficulty from test to test. For this reason, ISAT raw scores are transformed and reported as standard scores where numerically equivalent scores represent the same level of proficiency.

ISAT also uses two forms of the reading test at grades 5, 8, and 10. At each grade, two passages (and their associated items) are identical across the two forms and one passage is different. Because the two tests are not exactly equal in difficulty, scores on the two forms are statistically equated using the one-parameter (Rasch) model. The two forms were jointly calibrated, which places the difficulty of both sets of items on the same scale and makes proficiency estimates equivalent across test forms. IRT scaling is also used with the ISAT mathematics tests.

This approach places both sets of tests on a firm basis to meet future equating needs. Successive years' test forms, which will have different items, will be equated so that test scores will remain comparable across administrations. Each new test will contain a sufficient number of items that have been previously administered to provide a reliable and content-representative equating link. During calibration of the new tests, item difficulties for these linking items will be set to their historical values. By estimating values for the remaining items under this constraint, difficulty values for the remaining items will be automatically adjusted to the existing scale. The final step in the procedure is to apply equations that transform values on the proficiency scale to their corresponding scale score values. These equations are developed during the first year of testing and are then applied in each subsequent year.

ISAT reading and mathematics scores are reported on a standard score scale. Individual student scores on this scale range between 120 and 200, regardless of the characteristics of the raw score distribution. Each scale is defined by letting 160 represent the average estimated proficiency of the first-year test population. Every unit on the scale represents 1/15 of the standard deviation of proficiency scores for the first-year population. In other words, the first year mean and standard deviation of scale scores for each grade are 160 and 15. Results in subsequent years will be equated to the base-year scale. The scaling constants used to transform the Rasch proficiency estimates to the reporting scale are shown in Table 3.1.

**Table 3.1**
**ISAT Reading and Mathematics Scaling Constants**

| Reading | Slope | Intercept |
|---|---|---|
| Grade 3 | 12.6428 | 146.2066 |
| Grade 5 | 12.0100 | 144.7660 |
| Grade 8 | 11.2280 | 141.7730 |
| Grade 10 | 12.1470 | 140.1670 |
| | | |
| Mathematics | | |
| Grade 3 | 13.5122 | 147.6910 |
| Grade 5 | 14.9686 | 153.4644 |
| Grade 8 | 14.7578 | 146.7806 |
| Grade 10 | 13.0795 | 148.3945 |

The raw score that is initially derived from multiple-choice items in reading and mathematics has no particular meaning beyond the number of answers the student has answered correctly. Writing, on the other hand, uses criterion-referenced scales. Each point on these scales has a specific interpretation. For example, when readers evaluate the quality of a 3rd-grade persuasive essay's focus, they assign a score of 6 when the paper "sets its purpose in an introduction through either a general thematic introduction or a specific preview, maintains the position or logic throughout, addresses any previewed points, and provides an effective closing." They assign a score of 3 when the paper "lacks clarity, provides multiple positions with a unifying umbrella statement, contains responses that do not serve a persuasive purpose, or lacks sufficiency to demonstrate a developed focus." Transforming writing scores to another scale would lose the specific meanings attached to each score point. For this reason, the ISAT writing score is a simple summation of the features. Because of the importance of Integration, it is given double weight in the summation. This leads to a writing score that ranges from 6 to 32.

# 4. NATIONAL NORM COMPARISONS

The legislation that authorized the development of ISAT required that reports provide national comparative data as a secondary reference point for evaluating school improvement efforts. Since the costs of obtaining nationally representative samples of students for each test would be prohibitively expensive, that mandate has been met by administering a nationally standardized achievement test along with ISAT to a sample of Illinois students. The two score distributions are then compared to identify points on the ISAT scale that correspond to the 25th, 50th, and 75th percentile performance levels for the national sample.

ISAT uses the Ninth Edition of the Stanford Achievement Tests (SAT9) for purposes of determining Illinois students' relative standing within the national population. The specific levels/norms of each SAT9 test used were as follows:

Grade 3: Primary 3

Grade 5: Intermediate 2

Grade 8: Advanced 2

Grade 10: TASK 2

Equipercentile methodology was used to equate scores on the two tests. In equipercentile equating, scores on two tests are assumed to be equivalent if they have the same percentile rank. For example, the SAT9 score that cuts off 10% of the equating sample is assumed to represent a level of proficiency equal to the ISAT score that cuts off 10% of the equating sample, even though the scores themselves may be quite different numerically.

In order to conduct the equating process, ISAT and SAT9 results were matched by name to create a set of records in which each student had ISAT results and a corresponding SAT9 score. Frequency distributions of ISAT and SAT9 scale scores were then compiled. Each scale score on the ISAT was matched to the corresponding scale score on the Stanford test, based on the cumulative mid-percentile interval associated with each score.

For 1999, this process was conducted for all reading and mathematics tests. National norms for writing are not provided because no nationally standardized writing test has a sufficiently satisfactory match to the Illinois content specifications to be used for this purpose.

Table 4.1 summarizes results of these studies. It shows the sample sizes that were used for the equating, the average SAT9 national percentile for the samples, and the correlations between the two instruments. Table 4.2 presents the ISAT scale score cutoffs that define the *upper limits* of national quartile categories 1, 2, and 3. These are shown as score ranges for each national quarter. For example, scale scores of 120 to 147 on the 3rd-grade reading test define Q1, the quartile that represents the lowest 25% of student performance nationally. Note that although the scale score cutoffs remain the same from year to year, the percentage of students in each category need not remain constant.

**Table 4.1**
**Summary of 1999 ISAT–SAT9 National Norm Studies**

| READING | Grade 3 | Grade 5 | Grade 8 | Grade 10 |
|---|---|---|---|---|
| Sample Size | 998 | 1199 | 1083 | 1054 |
| Mean SAT9 Percentile | 55.9 | 53.2 | 58.2 | 53.7 |
| r between ISAT and SAT9 Scores | .837 | .805 | .799 | .679 |
| **MATHEMATICS** | **Grade 3** | **Grade 5** | **Grade 8** | **Grade 10** |
| Sample Size | 805 | 690 | 749 | 1357 |
| Mean SAT9 Percentile | 57.8 | 52.3 | 51.8 | 65.9 |
| r between ISAT and SAT9 Scores | .856 | .870 | .831 | .839 |

**Table 4.2**
**ISAT National Quarter Scale Score Cutoffs**

| READING | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 03 | 120-147 | 148-157 | 158-167 | 168-200 |
| 05 | 120-147 | 148-157 | 158-168 | 169-200 |
| 08 | 120-144 | 145-154 | 155-165 | 166-200 |
| 10 | 120-153 | 154-162 | 163-171 | 172-200 |
| **MATHEMATICS** | **Q1** | **Q2** | **Q3** | **Q4** |
| 03 | 120-145 | 146-155 | 156-166 | 167-200 |
| 05 | 120-146 | 147-156 | 157-166 | 167-200 |
| 08 | 120-144 | 145-154 | 155-164 | 165-200 |
| 10 | 120-146 | 147-154 | 155-164 | 165-200 |

The results of applying these cutoffs to the 1999 assessment data are shown in Table 4.3.

**Table 4.3**
**Percentages of Students by Grade and Learning Area Falling into Each National Quartile:**

READING

| Grade | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 3 | 22 | 22 | 25 | 32 |
| 5 | 21 | 23 | 27 | 28 |
| 8 | 15 | 22 | 30 | 33 |
| 10 | 30 | 25 | 23 | 21 |

MATHEMATICS

| Grade | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| 3 | 19 | 21 | 28 | 32 |
| 5 | 20 | 22 | 24 | 33 |
| 8 | 15 | 25 | 25 | 35 |
| 10 | 19 | 19 | 27 | 35 |

Note: Because of rounding, the percentages in each row may not total exactly to 100%.

# 5. RESULTS

Table 5.1 shows the percentages of students by performance level and by grade for reading. The percentage of students falling into the Exceeds category is highest at 5th grade. However, the percentage of students not meeting standards is also highest at 5th grade. Overall, the percentage of students meeting (or exceeding) standards is highest at 8th grade.

**Table 5.1**
**Percentages of Students by Grade Falling into Each Performance Level for ISAT Reading: 1999**

| Grade | Academic Warning | Below Standards | Meets Standards | Exceeds Standards |
|---|---|---|---|---|
| 3 | 8 | 31 | 44 | 17 |
| 5 | 1 | 38 | 37 | 24 |
| 8 | 1 | 27 | 54 | 18 |
| 10 | 5 | 25 | 55 | 15 |

Note: Because of rounding, the percentages in each row may not total exactly to 100%.

Table 5.2 provides additional information with respect to the reading test. It presents the average percent of items students answered correctly with respect to the standards sets that were previously described.

**Table 5.2**
**Reading Average Percent Correct by Standards Sets: 1999**

| | Set | | | | | |
|---|---|---|---|---|---|---|
| Grade | 1 | 2 | 3 | 4 | 5 | 6 |
| 03 | 74 | 63 | 71 | 70 | 69 | 67 |
| 05 | 74 | 72 | 76 | 71 | 65 | |
| 08 | 80 | 69 | 75 | 77 | 80 | |
| 10 | 76 | 78 | 81 | 75 | 76 | |

Table 5.3 shows the percentages of students by performance level and by grade for mathematics. Generally, the percentage of students meeting state standards is lower for mathematics than for reading. Grade 3 is an exception to that rule. The percentage of students falling into the Exceeds category is highest at 3rd grade and lowest at 5th grade.

**Table 5.3**
**Percentages of Students by Grade Falling into Each Performance Level for ISAT Mathematics: 1999**

| Grade | Academic Warning | Below Standards | Meets Standards | Exceeds Standards |
|---|---|---|---|---|
| 3 | 12 | 20 | 47 | 21 |
| 5 | 6 | 39 | 53 | 3 |
| 8 | 5 | 52 | 36 | 7 |
| 10 | 6 | 41 | 47 | 5 |

Note: Because of rounding, the percentages in each row may not total exactly to 100%.

Table 5.4 presents the average percent of items students answered correctly with respect to the mathematics standards sets that were previously described.

**Table 5.4**
**Mathematics Average Percent Correct by Standards Sets: 1999**

| | Set | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Grade | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 03 | 65 | 64 | | 55 | 63 | 62 | 72 | 58 |
| 05 | 56 | 59 | 52 | 59 | 66 | 47 | 52 | 55 |
| 08 | 51 | 51 | 53 | | 46 | 38 | 56 | 56 |
| 10 | 53 | 54 | 50 | | 50 | 53 | 64 | 64 |

Table 5.5 shows results for writing. With respect to the Exceeds category, there are wide differences between 5th grade, on the one hand, and 3rd grade and 8th grade on the other.

Table 5.6 summarizes results with respect to writing feature scores. Note that Conventions is scored on a two-point scale while all other features are scored on a six-point scale.

**Table 5.5**
**Percentages of Students by Grade Falling into Each Performance Level for ISAT Writing: 1999**

| Grade | Academic Warning | Below Standards | Meets Standards | Exceeds Standards |
|---|---|---|---|---|
| 3 | 9 | 35 | 50 | 6 |
| 5 | 2 | 23 | 52 | 23 |
| 8 | 5 | 36 | 56 | 3 |
| 10 | 6 | 28 | 54 | 12 |

Note: Because of rounding, the percentages in each row may not total exactly to 100%.

**Table 5.6**
**Mean Writing Feature Scores of Students by Prompt: 1999**

| Grade | Type | F | S | O | C | I |
|---|---|---|---|---|---|---|
| 03 | P | 4.52 | 3.67 | 3.69 | 1.94 | 3.89 |
| 03 | E | 4.49 | 3.72 | 3.69 | 1.94 | 3.90 |
| 03 | N | 3.92 | 3.93 | 3.77 | 1.90 | 3.87 |
| 05 | P | 4.64 | 4.20 | 4.22 | 1.94 | 4.31 |
| 05 | E | 4.64 | 4.15 | 4.16 | 1.95 | 4.27 |
| 05 | N | 4.29 | 4.38 | 4.21 | 1.95 | 4.28 |
| 08 | P | 3.93 | 3.74 | 3.72 | 1.90 | 3.77 |
| 08 | E | 3.95 | 3.74 | 3.69 | 1.90 | 3.76 |
| 08 | N | 4.03 | 4.11 | 3.93 | 1.87 | 4.03 |
| 10 | P | 4.10 | 3.89 | 3.89 | 1.94 | 3.92 |
| 10 | E | 4.11 | 3.87 | 3.86 | 1.95 | 3.90 |
| 10 | N | 4.35 | 4.36 | 4.28 | 1.94 | 4.35 |

Note: Prompt type: P = Persuasive; E = Expository; N = Narrative

# Part 2: The Illinois Goal Assessment Program

# 6. PURPOSE AND DESIGN OF THE IGAP TESTING PROGRAM

In February 1999, students in grades 4, 7, and 11 took Illinois Goal Assessment Program (IGAP) tests in science and social sciences. Statewide studies were also conducted in fine arts and physical development and health. IGAP measures the extent to which students are meeting the Illinois Learning Goals (1985). 1998-99 was the last school year for IGAP testing for science and social sciences. Beginning with the 1999-2000 school year, science and social science will be tested as ISAT areas.

This manual provides technical information about the 1999 test administration. It is an extension of earlier technical manuals. It focuses primarily on elements of the program that changed in 1999. Some material from previous manuals is repeated to provide a context for understanding current developments. Previous technical manuals provide additional technical details of the program and its history.

## General Procedures

Each test is designed to ensure that its results validly and fairly assess the extent to which schools and districts meet the state learning goals. The selection of items and assembly of each test is guided by a set of specifications. These specifications were developed by Illinois educators to help ensure that test content corresponds to the purposes, objectives, and skills framed by the state learning goals. The state learning goals represent Illinois' vision of and commitment to world-class education for its students and citizens.

Illinois teachers and administrators participate in all phases of the test development process: item writing, item selection, bias review, and test assembly. The State Board of Education convenes a series of advisory committees to ensure that test development is continually informed and guided by the recommendations of content authorities, measurement specialists, and practitioners. The following evaluation criteria are applied to all assessment material used in the Illinois program:

*Content.* Every item is screened for alignment with the Illinois Learning Standards, grade-level appropriateness, importance, and clarity. Incorrect choices (for multiple-choice items) are reviewed for plausibility. In tests other than reading, the complexity of the text of the questions is kept to the minimum necessary to state the problem.

*Difficulty.* Items are pilot tested on large samples of students prior to their inclusion in tests to develop a statistical profile for each item. Items that are too easy or too difficult and, therefore, provide little or no information are omitted.

*Precision.* Point-biserial (i.e., item-test) correlations evaluate the extent to which an item distinguishes between less proficient and more proficient students. Reviewers usually omit items with a point-biserial of less than .30 and select items with the highest point-biserial.

*Fairness.* Test items and forms undergo regular sensitivity reviews and statistical analyses to ensure that all material, meet fairness criteria with respect to the cultural and ethnic diversity of Illinois public schools.

The content of the tests changes each year so that knowledge of specific questions or assessment material does not spuriously inflate scores. Each new test is statistically equated to previous tests so that schools can accurately discern trends in performance across time. This topic is discussed extensively in Section 8.

ISBE takes several precautions to help ensure test security. Test materials shipped to schools are packaged and sealed. Each test booklet is bar-coded so that it can be accounted for. The administration of tests is standardized. A series of manuals provides guidance on security and other issues to the district testing coordinator, school testing coordinator, and classroom test administrator. After administration, all materials are removed from schools and returned to a central facility for processing and secure destruction of unneeded materials.

## Science

The state goals for learning in science specify that, as a result of their schooling, students will have a working knowledge of:

1 the concepts and basic vocabulary of biological, physical, and environmental sciences and their application to life and work in contemporary technological society;

2 the social and environmental implications and limitations of technological development;

3 the principles of scientific research and their application in simple research projects; and

4 the processes, techniques, methods, equipment, and available technology of science.

The science assessment consists of single-right-answer, multiple-choice items. Test items are distributed evenly across the four goal areas. In addition to the overall science scale score, which is based on 64 items, subtest scores are also reported. Goal scores for the schools, districts, and the state are also placed on the 0-500 scale on which IGAP scores are reported. A set of science pilot items and a set of health/physical development items used for conducting state studies bring the total number of items in each test to 80. The pilot items do not contribute to test scores.

The Productive Thinking Scale (PTS) is used to evaluate the quality of science items. It is hierarchical with respect to the production of knowledge and independent of an item's difficulty or grade. Four cognitive skills define the hierarchy of productive thinking in generating scientific knowledge. Each skill applies to both content (knowledge) and to process (research methods): (1) recall of conventions, whether names or norms; (2) reproduction of empirical facts or methodological tools and steps; (3) production of solutions

to problems or research designs; and (4) creation of new theories and methods. The PTS subdivides reproduction and production into secondary processes. Hence, the PTS comprises six levels of productive thinking on a scale from low level (recall of conventional uses) to high level (creation of new theory).

Based on estimates of the thought processes which most students must use to answer an item, each item is ranked as to the level of conceptual skill it requires. Items that provide a rough balance across the middle ranks are selected, and items at the level of vocabulary or rote memory are usually omitted.

Items are also examined to determine whether there is a reasonable distribution within tests of items among major learning areas: earth science, physical science, and life science.

## Social Sciences

Social sciences provide students with an understanding of themselves and of society, prepare them for citizenship in a democracy, and give them the basics for understanding the complexities of the world community. Social sciences include anthropology, economics, geography, government, history, political science, psychology, and sociology. Five goals in social sciences define what students should know and be able to do at each grade:

1 understand and analyze comparative political and economic systems, with an emphasis on the political and economic systems of the United States;

2 understand and analyze events, trends, personalities, and movements shaping the history of the world, the United States, and Illinois;

3 demonstrate knowledge of the basic concepts of the social sciences and how these help to interpret human behavior;

4 demonstrate a knowledge of world geography with an emphasis on the United States; and

5 apply the skills and knowledge gained in the social sciences to decision making in life situations.

This delineation of what students should know and be able to do defines the five areas assessed by the social sciences tests: Political and Economic Systems, History, Basic Concepts of Social Science, Geography, and Application and Decision Making.

The IGAP social sciences assessment consists of single-right-answer, multiple-choice items. Test items are distributed approximately evenly across the five goal areas. In addition to the overall social sciences scale score, which is based on the complete set of 75 items, subtest scores are also reported. Goal scores for the schools, districts, and the state are also placed on the 0-500 scale. A set of fine arts items used for conducting state studies brings the total number of items in each test to 81.

# 7. RELIABILITY

## Internal Consistency of Overall Scores

Because the items used in achievement tests represent only a relatively small sample from a much larger domain of items, the consistency of test scores across items is of particular interest. That is, how precisely will tests rank students if different sets of items from the same domain are used? Unless the rankings are very similar, it is difficult or impossible to make educationally sound decisions on the basis of test scores. This characteristic of test scores is most commonly referred to as *internal consistency*. Table 7.1 presents internal consistency values (coefficient alpha) for each of the IGAP tests administered in the 1999 assessment. The samples represented 1/nth selections from the total population. All of the tests show a high level of reliability. These coefficients are comparable to those reported for commercially available, nationally standardized achievement tests.

**Table 7.1**
**1999 Reliability Estimates**

| Grade | Science | Social Sciences |
|---|---|---|
| 04 | .92 | .93 |
| 07 | .91 | .93 |
| 11 | .91 | .91 |

Note: Sample sizes on which these coefficients are based are as follows:

Science: 4 (15,912), 7 (15,805), 11 (15,739)
Social Sciences: 4 (15,846), 7 (15,749), 11 (15,729)

## Reliability of the Performance Category Decisions

Students' scores are reported relative to sets of score cutoffs that were developed for each learning area and each grade. The 1999 IGAP scores are reported relative to three categories: Below Standards, Meets Standards, and Exceeds Standards. As noted earlier in Section 2, these reliabilities are not equivalent to the test reliabilities.

Table 7.2 shows the 1999 values for $p$, $k$, and $p_{miss}$ for the IGAP tests. A comparison of values shown in Table 2.6 shows that, overall, decisions appear to be more consistent for ISAT than IGAP tests. This likely reflects the fact that the ISAT tests are somewhat longer than the IGAP tests.

Note that the $p$ and $k$ values reported in Table 7.2 are calculated for the complete test population. Values for other test populations (e.g., IEP students alone, non-IEP students alone) may differ.

**Table 7.2**
**Reliability of Student Performance Decisions Based on 1999 Test Scores**

| | | Below Standards/Meets Standards | | | Meets Standards/Exceeds Standards | | |
|---|---|---|---|---|---|---|---|
| Area | Grade | p | kappa | $p_{miss}$ | p | kappa | $p_{miss}$ |
| Science | 4 | .932 | .666 | .068 | .874 | .705 | .126 |
| | 7 | .910 | .683 | .090 | .876 | .699 | .124 |
| | 11 | .900 | .685 | .100 | .902 | .691 | .098 |
| Social Sciences | 4 | .926 | .779 | .074 | .912 | .787 | .088 |
| | 7 | .936 | .774 | .064 | .912 | .794 | .088 |
| | 11 | .926 | .662 | .074 | .896 | .689 | .104 |
| AVERAGE | | .922 | .708 | .078 | .895 | .728 | .105 |

# 8. IGAP EQUATING PROCEDURES

Without equating, each administration of a test with different items would lead to a new reporting scale, independent of that used previously. It would still be possible to measure relative performance, but it would not be possible to indicate growth across years for schools, districts, or the state. The equating process makes longitudinal comparisons possible.

Different procedures have been used at various times in the history of IGAP to accommodate changing needs of each learning area. In 1993, however, ISBE implemented a plan to bring the scaling/equating designs and procedures into alignment across areas, where a common methodology was appropriate. Based on recommendations from the Assessment Advisory Committee, IGAP tests employ IRT true score equating using the one-parameter (Rasch) model to place each year's results onto the reporting scale.

The equating procedures may be summarized as follows. Each test contains a sufficient number of items that have been previously administered to provide a reliable and content-representative equating link. During calibration of the new tests, item difficulties for these linking items are set to their historical values. By estimating values for the remaining items under this constraint, difficulty values for the remaining items are expressed on the existing scale. That is, the proficiency (theta) scale that results from the constrained calibration run is equated to the existing scale. The final step in the procedure is to apply equations that transform values on the proficiency scale to their corresponding IGAP scale score values. These equations were originally developed during the first year of equating and are then applied in each subsequent year of equating.

The logic of the equating procedure rests on certain assumptions. The most important is that the items used for linking stay the same in the two tests. During the assembly of tests, items that will be used for equating are placed exactly at or very near the location in the booklet where they previously appeared (i.e., item 23 in 1998 is also item 23 in 1999) to minimize effects from positional differences. Differences between the anchored difficulties and the best-fit values are examined to ensure that no unusually large differences exist that would strain the equivalence assumption. The difference in average item difficulties between the anchored and unanchored calibration runs is called the equating constant. Ideally, this value should be relatively small.

The equating analyses are conducted on samples of approximately 16,000 drawn from the total test population. A 1/nth selection results in a sample that has characteristics essentially identical with that of the total population.

Results of the Rasch equatings are shown as follows: science: Tables 8.1 through 8.3; social sciences: Tables 8.4 through 8.6. The format of these tables is identical. Column 1 shows the item position in the 1999 test. Column 2 shows the Rasch difficulties resulting from an unanchored (unconstrained) calibration of the 1999 test. Column 3 shows the item position in the 1998 test. Column 4 shows the difficulty values at which the 1999 item difficulties were anchored. The last column in each table shows the learning goal associated with each item. As noted earlier, the *equating constant* shown below each table is the mean difference in difficulty values and represents the amount that must be added to the 1999 values to

place them on the original scale. When the proficiency scale is recomputed from the adjusted difficulties and multiplied by the scaling constants that relate proficiency to scale scores, the result is a table that converts 1999 raw scores to scale scores through the linked proficiency scales.

**Table 8.1**
**Item Difficulty Parameters for Science Items Common to 1999 and 1998: Grade 4**

| 1999 Test | | 1998 Test | | |
|---|---|---|---|---|
| Position | Difficulty | Position | Difficulty | Goal |
| 1 | -0.10 | 1 | 0.06 | 1 |
| 2 | -0.61 | 2 | -0.37 | 1 |
| 3 | -0.31 | 3 | -0.10 | 1 |
| 4 | -0.86 | 4 | -0.56 | 1 |
| 5 | 0.07 | 5 | 0.43 | 1 |
| 8 | -0.56 | 8 | -0.19 | 1 |
| 9 | 0.68 | 9 | 0.93 | 2 |
| 10 | -0.77 | 10 | -0.46 | 2 |
| 11 | -0.22 | 11 | 0.02 | 2 |
| 12 | 0.33 | 12 | 0.62 | 2 |
| 13 | -0.86 | 13 | -0.48 | 2 |
| 14 | 1.16 | 14 | 1.62 | 2 |
| 15 | -0.20 | 15 | 0.43 | 2 |
| 16 | -0.86 | 16 | -0.62 | 2 |
| 17 | -0.36 | 17 | -0.33 | 3 |
| 18 | 0.14 | 18 | 0.42 | 3 |
| 19 | -0.79 | 19 | -0.59 | 3 |
| 20 | -0.39 | 20 | -0.08 | 3 |
| 21 | -0.43 | 21 | -0.25 | 3 |
| 22 | -0.01 | 22 | 0.15 | 3 |
| 25 | -0.01 | 25 | 0.25 | 4 |
| 29 | 0.90 | 29 | 1.17 | 4 |
| 30 | -0.26 | 30 | 0.17 | 4 |
| 31 | 1.18 | 31 | 1.28 | 4 |
| 42 | 0.63 | 42 | 0.92 | 4 |
| 46 | 0.24 | 46 | 0.79 | 4 |
| 47 | 0.12 | 47 | 0.42 | 4 |
| 48 | 0.17 | 48 | 0.21 | 4 |
| 49 | 0.26 | 49 | 0.63 | 3 |
| 50 | 0.02 | 50 | 0.53 | 3 |
| 52 | -0.23 | 52 | -0.04 | 3 |
| 53 | 0.05 | 53 | 0.34 | 3 |
| 54 | 0.20 | 54 | 0.41 | 3 |
| 55 | -0.93 | 55 | -0.41 | 3 |
| 57 | 0.90 | 57 | 1.32 | 2 |
| 58 | -0.10 | 58 | 0.27 | 2 |
| 59 | 0.12 | 59 | 0.46 | 2 |
| 60 | -0.34 | 60 | -0.14 | 2 |
| 61 | -0.37 | 61 | 0.18 | 2 |
| 62 | 0.25 | 62 | 0.57 | 2 |
| 63 | -0.31 | 63 | 0.11 | 2 |
| 64 | -0.52 | 64 | -0.22 | 1 |
| 65 | 0.16 | 65 | 0.40 | 1 |
| 66 | -0.48 | 66 | -0.17 | 1 |
| 67 | 0.46 | 67 | 0.78 | 1 |
| 71 | 0.29 | 71 | 0.68 | 1 |
| 72 | 0.75 | 72 | 1.20 | 1 |

Equating Constant: .31

**Table 8.2**
**Item Difficulty Parameters for Science Items Common to 1999 and 1998: Grade 7**

| 1999 Test | | 1998 Test | | |
|---|---|---|---|---|
| Position | Difficulty | Position | Difficulty | Goal |
| 2 | -0.30 | 2 | -0.30 | 1 |
| 3 | -0.92 | 3 | -0.75 | 1 |
| 4 | -1.42 | 4 | -1.52 | 1 |
| 5 | 0.13 | 5 | -0.07 | 1 |
| 7 | -1.30 | 7 | -1.39 | 1 |
| 8 | 0.65 | 8 | 0.58 | 1 |
| 9 | 0.01 | 9 | -0.06 | 2 |
| 10 | -1.15 | 10 | -1.15 | 2 |
| 11 | -0.47 | 11 | -0.52 | 2 |
| 13 | -1.19 | 13 | -1.13 | 2 |
| 14 | -1.60 | 14 | -1.56 | 2 |
| 15 | 0.49 | 15 | 0.40 | 2 |
| 16 | 0.27 | 16 | 0.06 | 2 |
| 18 | -0.84 | 18 | -0.89 | 3 |
| 20 | 0.16 | 20 | 0.04 | 3 |
| 21 | 0.46 | 21 | 0.31 | 3 |
| 23 | -0.65 | 23 | -0.70 | 3 |
| 25 | -0.15 | 25 | -0.26 | 4 |
| 27 | 0.26 | 27 | -0.10 | 4 |
| 28 | 0.13 | 28 | 0.02 | 4 |
| 29 | -0.41 | 29 | -0.56 | 4 |
| 31 | 0.40 | 31 | -0.11 | 4 |
| 32 | 0.92 | 32 | 0.79 | 4 |
| 41 | 0.54 | 41 | 0.57 | 4 |
| 42 | 1.34 | 42 | 1.30 | 4 |
| 43 | 0.79 | 43 | 0.76 | 4 |
| 44 | -1.27 | 44 | -1.32 | 4 |
| 47 | -0.97 | 47 | -1.28 | 4 |
| 48 | -0.27 | 48 | -0.26 | 4 |
| 53 | -0.92 | 53 | -0.97 | 3 |
| 54 | -0.10 | 54 | 0.00 | 3 |
| 56 | 0.16 | 56 | 0.28 | 3 |
| 57 | -0.86 | 57 | -0.95 | 2 |
| 58 | 0.46 | 58 | 0.38 | 2 |
| 61 | -0.19 | 61 | -0.27 | 2 |
| 62 | -0.06 | 62 | 0.20 | 2 |
| 63 | 0.54 | 63 | 0.44 | 2 |
| 64 | -0.26 | 64 | -0.41 | 2 |
| 65 | 0.04 | 65 | 0.44 | 1 |
| 66 | 0.29 | 66 | 0.13 | 1 |
| 70 | 0.38 | 70 | 0.19 | 1 |
| 71 | 0.41 | 71 | 0.30 | 1 |
| 72 | 0.45 | 72 | 0.32 | 1 |

Equating Constant: -.07

**Table 8.3**
**Item Difficulty Parameters for Science Items Common to 1999 and 1998: Grade 11**

| 1999 Test | | 1998 Test | | |
|---|---|---|---|---|
| Position | Difficulty | Position | Difficulty | Goal |
| 2 | 1.12 | 2 | 1.50 | 1 |
| 3 | 1.04 | 3 | 1.13 | 1 |
| 6 | 0.24 | 6 | 0.35 | 1 |
| 7 | 0.76 | 7 | 0.83 | 1 |
| 8 | 1.27 | 8 | 1.62 | 2 |
| 9 | -0.14 | 9 | 0.06 | 3 |
| 11 | -0.26 | 11 | -0.02 | 2 |
| 16 | 0.35 | 16 | 0.45 | 2 |
| 17 | 0.27 | 17 | 0.62 | 3 |
| 19 | -0.90 | 19 | -0.80 | 3 |
| 20 | -0.05 | 20 | 0.17 | 3 |
| 21 | -0.41 | 21 | -0.06 | 3 |
| 23 | -0.73 | 23 | -0.51 | 3 |
| 24 | -0.40 | 24 | -0.07 | 3 |
| 26 | 0.13 | 26 | 0.36 | 4 |
| 27 | 1.04 | 27 | 1.26 | 4 |
| 29 | 0.27 | 29 | 0.36 | 4 |
| 30 | -0.38 | 30 | -0.17 | 4 |
| 41 | -0.01 | 41 | 0.09 | 4 |
| 44 | 0.40 | 44 | 0.92 | 4 |
| 45 | -0.61 | 45 | -0.06 | 4 |
| 46 | -1.07 | 46 | -0.83 | 4 |
| 48 | 0.77 | 48 | 0.93 | 4 |
| 49 | 0.40 | 49 | 0.54 | 3 |
| 50 | -0.22 | 50 | 0.04 | 3 |
| 51 | 0.11 | 51 | 0.48 | 3 |
| 52 | -0.07 | 52 | 0.17 | 3 |
| 53 | -0.02 | 53 | 0.27 | 3 |
| 57 | 0.19 | 57 | 0.35 | 3 |
| 58 | 0.06 | 58 | 0.38 | 2 |
| 59 | 0.55 | 59 | 0.69 | 2 |
| 60 | -0.23 | 60 | 0.09 | 2 |
| 61 | 0.71 | 61 | 0.80 | 2 |
| 62 | 0.67 | 62 | 0.91 | 2 |
| 63 | -0.38 | 63 | -0.21 | 2 |
| 65 | 0.03 | 65 | 0.39 | 1 |
| 67 | -0.06 | 67 | 0.14 | 1 |
| 68 | -0.88 | 68 | -0.57 | 1 |
| 70 | -0.44 | 70 | -0.17 | 1 |
| 71 | -0.27 | 71 | -0.20 | 1 |

Equating Constant: .23

**Table 8.4**
**Item Difficulty Parameters for Social Sciences Items Common to 1999 and 1998: Grade 4**

| 1999 Test | | 1998 Test | | |
|---|---|---|---|---|
| Position | Difficulty | Position | Difficulty | Goal |
| 4 | -0.63 | 4 | -0.60 | 1 |
| 5 | -0.22 | 5 | -0.18 | 5 |
| 7 | -0.24 | 7 | -0.32 | 1 |
| 8 | 0.21 | 8 | 0.20 | 1 |
| 10 | 0.15 | 10 | -0.12 | 2 |
| 12 | -0.55 | 12 | -0.46 | 2 |
| 13 | 0.26 | 13 | 0.18 | 2 |
| 18 | 0.33 | 17 | 0.47 | 2 |
| 19 | 0.73 | 18 | 0.68 | 3 |
| 22 | 0.86 | 21 | 0.57 | 4 |
| 23 | -0.06 | 22 | -0.34 | 3 |
| 25 | 0.91 | 24 | 0.89 | 4 |
| 26 | -0.83 | 25 | -1.12 | 4 |
| 28 | 0.13 | 27 | -0.09 | 4 |
| 34 | 0.17 | 33 | 0.03 | 5 |
| 35 | 0.48 | 34 | 0.45 | 5 |
| 38 | 0.02 | 38 | -0.27 | 5 |
| 39 | 0.44 | 39 | 0.11 | 5 |
| 40 | 0.88 | 40 | 0.43 | 5 |
| 41 | -0.95 | 41 | -1.09 | 5 |
| 42 | 0.83 | 42 | 0.55 | 5 |
| 43 | -0.56 | 43 | -0.43 | 5 |
| 44 | -1.63 | 44 | -1.53 | 5 |
| 45 | -0.64 | 47 | -0.57 | 4 |
| 46 | -1.33 | 48 | -1.34 | 4 |
| 47 | -0.13 | 49 | 0.03 | 4 |
| 48 | -0.74 | 45 | -0.54 | 4 |
| 55 | 0.29 | 56 | 0.29 | 2 |
| 56 | -0.09 | 57 | -0.14 | 2 |
| 57 | 0.01 | 58 | -0.06 | 2 |
| 58 | 0.19 | 59 | 0.11 | 2 |
| 59 | 0.10 | 60 | 0.04 | 2 |
| 60 | -0.67 | 61 | -0.75 | 2 |
| 61 | -0.89 | 62 | -0.65 | 2 |
| 63 | -0.35 | 63 | -0.49 | 1 |
| 65 | 0.24 | 65 | -0.15 | 1 |
| 66 | -0.65 | 68 | -1.29 | 5 |
| 67 | -0.11 | 67 | -0.40 | 1 |
| 68 | 0.22 | 66 | 0.46 | 1 |
| 69 | 0.69 | 69 | 0.61 | 1 |
| 70 | 0.00 | 70 | -0.16 | 1 |

Equating Constant: -.09

**Table 8.5**
**Item Difficulty Parameters for Social Sciences Items Common to 1999 and 1998: Grade 7**

| 1999 Test | | 1998 Test | | |
|---|---|---|---|---|
| Position | Difficulty | Position | Difficulty | Goal |
| 1 | -0.46 | 1 | -0.70 | 1 |
| 2 | 0.07 | 2 | 0.08 | 1 |
| 3 | 1.17 | 3 | 1.12 | 1 |
| 4 | -0.12 | 4 | -0.14 | 1 |
| 5 | 0.89 | 5 | 0.74 | 1 |
| 6 | -1.35 | 6 | -1.27 | 1 |
| 7 | -0.67 | 7 | -0.74 | 1 |
| 8 | 0.23 | 8 | 0.04 | 1 |
| 9 | 0.69 | 9 | 0.61 | 1 |
| 11 | -0.79 | 11 | -0.99 | 5 |
| 12 | 0.28 | 12 | -0.13 | 2 |
| 13 | -2.30 | 13 | -2.60 | 2 |
| 14 | -0.86 | 14 | -1.04 | 2 |
| 15 | 0.52 | 15 | 0.17 | 2 |
| 16 | -1.70 | 16 | -1.69 | 2 |
| 17 | -0.32 | 17 | -0.58 | 2 |
| 18 | -0.56 | 18 | -0.77 | 2 |
| 19 | 0.32 | 19 | 0.17 | 2 |
| 25 | 0.21 | 21 | -0.16 | 3 |
| 26 | -0.03 | 22 | -0.11 | 3 |
| 27 | 0.33 | 25 | 0.18 | 4 |
| 28 | 0.44 | 26 | 0.32 | 4 |
| 29 | 0.09 | 27 | 0.16 | 4 |
| 35 | 1.22 | 31 | 0.81 | 4 |
| 36 | 0.28 | 33 | 0.05 | 5 |
| 37 | 1.04 | 36 | 0.71 | 5 |
| 39 | -0.19 | 37 | -0.41 | 5 |
| 40 | -0.46 | 38 | -0.51 | 3 |
| 41 | -0.55 | 41 | -0.79 | 5 |
| 42 | 0.49 | 42 | 0.46 | 5 |
| 49 | 1.05 | 46 | 1.27 | 4 |
| 50 | -0.45 | 51 | -0.70 | 4 |
| 51 | 1.14 | 52 | 1.10 | 4 |
| 60 | 0.94 | 63 | 1.26 | 2 |
| 61 | 0.44 | 64 | 0.42 | 2 |
| 68 | 1.19 | 68 | 1.15 | 1 |
| 69 | 0.25 | 69 | 0.20 | 1 |
| 70 | 1.19 | 70 | 0.54 | 1 |

Equating Constant: -.14

**Table 8.6**
**Item Difficulty Parameters for Social Sciences Items Common to 1999 and 1998: Grade 11**

| 1999 Test | | 1998 Test | | |
|---|---|---|---|---|
| Position | Difficulty | Position | Difficulty | Goal |
| 1 | -1.35 | 1 | -1.30 | 1 |
| 2 | -0.93 | 2 | -0.88 | 1 |
| 3 | -2.26 | 3 | -2.13 | 1 |
| 5 | 0.51 | 5 | 0.44 | 1 |
| 6 | 0.28 | 6 | 0.67 | 1 |
| 7 | 0.26 | 7 | 0.28 | 1 |
| 8 | 0.85 | 8 | 0.95 | 1 |
| 10 | 0.96 | 9 | 0.95 | 2 |
| 11 | 0.50 | 10 | 0.61 | 2 |
| 12 | -0.79 | 11 | -0.57 | 2 |
| 13 | 1.08 | 13 | 0.99 | 2 |
| 14 | 0.51 | 14 | 0.50 | 2 |
| 22 | -2.70 | 23 | -2.68 | 4 |
| 23 | 0.20 | 24 | 0.33 | 4 |
| 24 | -1.39 | 25 | -1.48 | 4 |
| 25 | -0.30 | 26 | -0.41 | 4 |
| 28 | 0.25 | 29 | 0.23 | 5 |
| 29 | 0.22 | 30 | 0.28 | 5 |
| 30 | 0.94 | 31 | 1.14 | 2 |
| 31 | 0.09 | 32 | -0.13 | 5 |
| 32 | -0.54 | 33 | -0.77 | 5 |
| 34 | 0.73 | 36 | 0.83 | 5 |
| 35 | -0.63 | 38 | -0.74 | 5 |
| 36 | -0.14 | 39 | 0.03 | 5 |
| 37 | 0.16 | 40 | 0.24 | 5 |
| 45 | 0.68 | 45 | 0.57 | 4 |
| 46 | -0.19 | 46 | -0.05 | 4 |
| 47 | -0.51 | 47 | -0.52 | 4 |
| 48 | 1.62 | 48 | 1.72 | 4 |
| 55 | 0.06 | 56 | 0.06 | 2 |
| 56 | 0.11 | 57 | -0.03 | 2 |
| 57 | 0.53 | 58 | 0.56 | 2 |
| 58 | 0.25 | 59 | 0.36 | 2 |
| 59 | 0.71 | 60 | 0.73 | 2 |
| 60 | 0.32 | 63 | 0.37 | 2 |
| 67 | -0.83 | 67 | -0.74 | 1 |
| 68 | -0.08 | 68 | -0.22 | 1 |
| 69 | -0.13 | 69 | -0.10 | 1 |
| 70 | 0.11 | 70 | 0.09 | 1 |

Equating Constant: .03

## Goal Scores (Science, Social Sciences)

The equating of IGAP goal scores followed a similar procedure. Difficulty values from the full test calibration were used to generate proficiency scales for each goal score to be equated. Then, a parallel process used with the overall scale score was followed to produce the raw score to scale score conversion table.

Goal scores are reported on the same scale as the overall score. That is, the same equation that transforms overall proficiency estimates is applied to proficiency estimates for each goal score. This produces a desirable situation in which goal scores are reported on the same scale and differences among them can be meaningfully interpreted at the aggregate level. For various theoretical and computational reasons, averaging goal scores will not necessarily reproduce the overall scale score. However, they are comparable among themselves.

# 9. RESULTS

State average science scores since testing began are shown in Figure 9.1. Legislation changed the grades at which science tests were administered in 1992 and 1993, which explains the different start dates across grades. In 1992, the first year of science assessment, the 11th-grade science mean was 250. This increased to 257 in 1993, where it remained essentially unchanged until 1997, when it increased to 260. It climbed slightly again in 1998 to 263 and then fell back to 261 in 1999. Scores at 4th and 7th grades also dropped from 1998 to 1999. Grade seven jumped seven points from 1997 to 1998 and dropped six points from 1998 to 1999. The decline was more pronounced at 4th grade. Table 9.1 summarizes test results at the goal level.

**Figure 9.1**
**State Means for Science: 1992-1999**

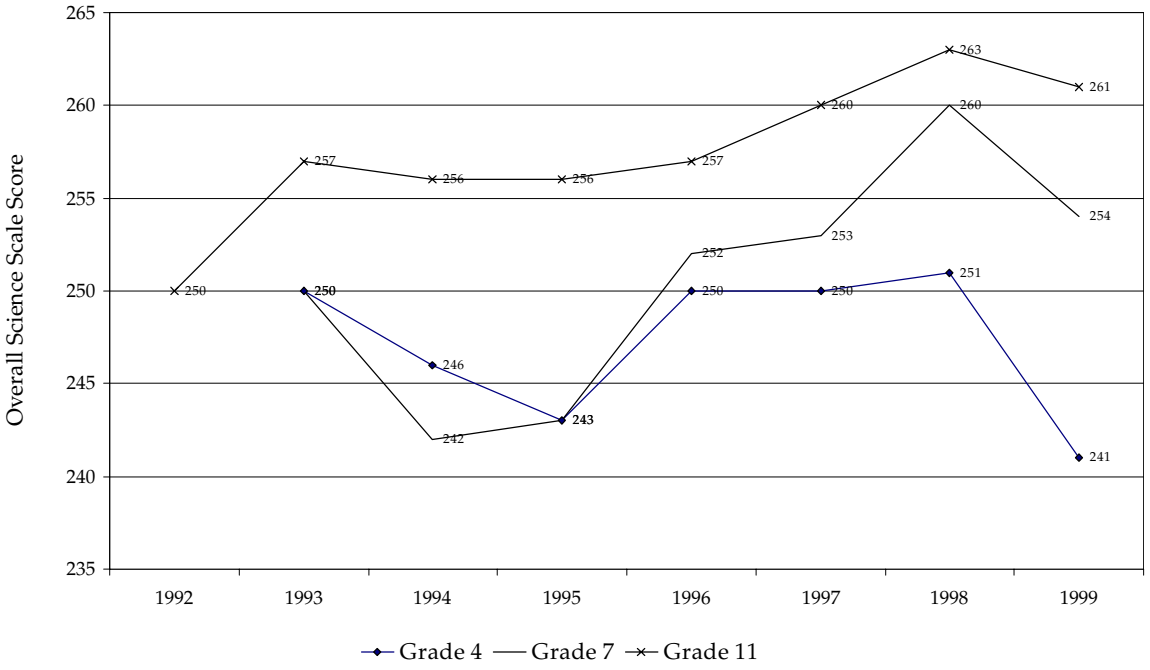


**Table 9.1**
**Science Goal State Means: 1992-1999**

| | | Goal | | | |
|---|---|---|---|---|---|
| Grade | Year | 1 | 2 | 3 | 4 |
| 04 | 1993 | 250 | 250 | 250 | 250 |
| | 1994 | 248 | 252 | 249 | 250 |
| | 1995 | 247 | 244 | 251 | 241 |
| | 1996 | 254 | 247 | 258 | 255 |
| | 1997 | 251 | 246 | 259 | 249 |
| | 1998 | 258 | 256 | 260 | 246 |
| | 1999 | 238 | 245 | 244 | 233 |
| 07 | 1993 | 250 | 250 | 250 | 250 |
| | 1994 | 243 | 247 | 245 | 243 |
| | 1995 | 245 | 241 | 251 | 245 |
| | 1996 | 253 | 256 | 260 | 251 |
| | 1997 | 251 | 261 | 257 | 253 |
| | 1998 | 259 | 267 | 268 | 264 |
| | 1999 | 248 | 252 | 257 | 248 |
| 11 | 1992 | 250 | 250 | 250 | 250 |
| | 1993 | 257 | 257 | 260 | 254 |
| | 1994 | 255 | 253 | 262 | 263 |
| | 1995 | 255 | 254 | 261 | 259 |
| | 1996 | 257 | 254 | 262 | 265 |
| | 1997 | 257 | 263 | 261 | 265 |
| | 1998 | 262 | 269 | 267 | 266 |
| | 1999 | 257 | 256 | 260 | 261 |

All pupil testing in social sciences began in 1993. For the first year of assessment the state means for the overall test score and goal area scores were set to 250. State average social sciences scores since 1993 are shown in Figure 9.2. For the first four years of the program, there was little evidence of a discernible trend in scores over time. During the next two years, scores for the elementary grades showed significant improvement, and grade 11 students showed some improvement. In 1999, however, scores returned to their earlier levels. Table 9.2 summarizes test results at the goal level.

**Figure 9.2**
**State Means for Social Sciences: 1993-1999**

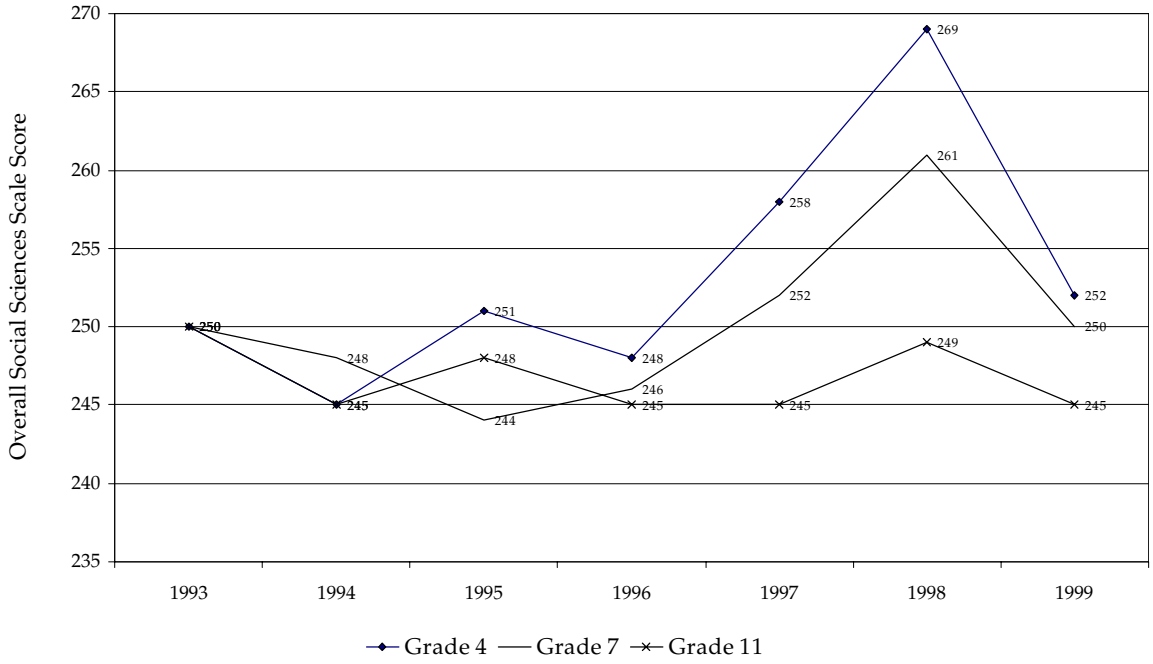


**Table 9.2**
**Social Sciences Goal State Means: 1993-1999**

| | | Goal | | | | |
|---|---|---|---|---|---|---|
| Grade | Year | 1 | 2 | 3 | 4 | 5 |
| 04 | 1993 | 250 | 250 | 250 | 250 | 250 |
| | 1994 | 259 | 248 | 250 | 246 | 252 |
| | 1995 | 259 | 254 | 254 | 256 | 256 |
| | 1996 | 259 | 245 | 246 | 256 | 244 |
| | 1997 | 259 | 254 | 266 | 262 | 262 |
| | 1998 | 272 | 274 | 274 | 269 | 274 |
| | 1999 | 249 | 253 | 249 | 252 | 245 |
| 07 | 1993 | 250 | 250 | 250 | 250 | 250 |
| | 1994 | 246 | 246 | 249 | 264 | 248 |
| | 1995 | 248 | 239 | 243 | 253 | 246 |
| | 1996 | 251 | 255 | 249 | 245 | 252 |
| | 1997 | 251 | 257 | 267 | 249 | 267 |
| | 1998 | 257 | 259 | 265 | 263 | 269 |
| | 1999 | 243 | 244 | 239 | 241 | 245 |
| 11 | 1993 | 250 | 250 | 250 | 250 | 250 |
| | 1994 | 242 | 249 | 250 | 253 | 246 |
| | 1995 | 241 | 253 | 255 | 256 | 246 |
| | 1996 | 244 | 252 | 248 | 252 | 248 |
| | 1997 | 243 | 254 | 249 | 249 | 247 |
| | 1998 | 250 | 248 | 258 | 252 | 251 |
| | 1999 | 240 | 240 | 243 | 248 | 238 |

## IGAP Student Performance Standards

With respect to performance standards, the data tell a similar but slightly different story than the scale score means. The IGAP scales were each established on the basis of different samples of students; scores are not directly comparable across grades (e.g., a 250 on the 6th-grade reading test isn't directly comparable in terms of student proficiency to a 250 on the 6th-grade mathematics test). Each scale compares a student's performance to a different norm group. In each case, the norm group was the group of students tested the first year an IGAP test was administered statewide.

The performance categories provide a second type of comparison, which is criterion-referenced rather than norm-referenced. Each performance category corresponds to a carefully defined level of performance. Thus, the performance categories provide a portrait of performance against a fixed set of standards rather than against a normative group. Results for science are presented in Table 9.3. Social sciences results appear in Table 9.4.

**Table 9.3**
**Percentages of Students by Grade Falling into Each Performance Level for IGAP Science: 1993-1999**

| Grade | Year | Below Standards | Meets Standards | Exceeds Standards |
|---|---|---|---|---|
| 4 | 1993 | 13 | 50 | 38 |
| | 1994 | 17 | 46 | 36 |
| | 1995 | 11 | 54 | 34 |
| | 1996 | 14 | 47 | 40 |
| | 1997 | 11 | 52 | 37 |
| | 1998 | 10 | 54 | 36 |
| | 1999 | 12 | 56 | 32 |
| 7 | 1993 | 22 | 45 | 34 |
| | 1994 | 21 | 51 | 28 |
| | 1995 | 20 | 50 | 30 |
| | 1996 | 18 | 49 | 32 |
| | 1997 | 16 | 53 | 30 |
| | 1998 | 12 | 54 | 34 |
| | 1999 | 19 | 49 | 32 |
| 11 | 1993 | 23 | 54 | 23 |
| | 1994 | 25 | 53 | 22 |
| | 1995 | 25 | 55 | 21 |
| | 1996 | 23 | 55 | 23 |
| | 1997 | 20 | 58 | 21 |
| | 1998 | 19 | 58 | 23 |
| | 1999 | 22 | 56 | 22 |

Note: Because of rounding, the percentages in each row may not total exactly to 100%.

**Table 9.4**
**Percentages of Students by Grade Falling into Each Performance Level for IGAP Social Sciences: 1993-1999**

| Grade | Year | Below Standards | Meets Standards | Exceeds Standards |
|---|---|---|---|---|
| 4 | 1993 | 19 | 51 | 30 |
| | 1994 | 22 | 49 | 29 |
| | 1995 | 19 | 48 | 33 |
| | 1996 | 19 | 53 | 28 |
| | 1997 | 19 | 47 | 34 |
| | 1998 | 17 | 46 | 37 |
| | 1999 | 22 | 49 | 29 |
| 7 | 1993 | 13 | 56 | 31 |
| | 1994 | 13 | 56 | 30 |
| | 1995 | 12 | 58 | 30 |
| | 1996 | 14 | 55 | 31 |
| | 1997 | 16 | 49 | 35 |
| | 1998 | 15 | 48 | 37 |
| | 1999 | 19 | 49 | 32 |
| 11 | 1993 | 10 | 66 | 24 |
| | 1994 | 14 | 62 | 25 |
| | 1995 | 12 | 62 | 25 |
| | 1996 | 11 | 66 | 23 |
| | 1997 | 12 | 64 | 24 |
| | 1998 | 11 | 66 | 23 |
| | 1999 | 13 | 66 | 21 |

Note: Because of rounding, the percentages in each row may not total exactly to 100%.

## National Quarters

IGAP uses the Stanford Achievement Tests (SAT) for purposes of determining Illinois students' relative standing within the national population. The studies linking the tests were conducted in 1993 (1992 for the 11th grade science test).

Prior to 1997, IGAP used the eighth edition of the SAT for purposes of providing national norms. In 1997, IGAP switched to the ninth edition (SAT9), which was standardized in 1995. The 1999 reports show results for both sets of cutoffs. Results for the earlier edition are presented to facilitate comparisons with previous years' reports.

Results for the 1998 assessment and for all previous years that IGAP has been administered at the same grade are shown in Table 9.5. The break between results for 1998 and previous years reflects the change in SAT editions (and the resulting cutoff scores) used to provide national comparisons.

**Table 9.5**
**Percentages of Students by Grade and Learning Area Falling into Each National Quartile: 1992-1999**

SCIENCE

| Grade | Year | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|---|
| 4 | 1993 | 19 | 20 | 24 | 38 |
| | 1994 | 23 | 20 | 20 | 36 |
| | 1995 | 22 | 24 | 23 | 32 |
| | 1996 | 22 | 20 | 22 | 37 |
| | 1997 | 28 | 25 | 25 | 22 |
| | 1998 | 26 | 29 | 23 | 22 |
| | 1999 | 32 | 29 | 21 | 18 |
| 7 | 1993 | 22 | 22 | 23 | 34 |
| | 1994 | 24 | 24 | 24 | 28 |
| | 1995 | 22 | 25 | 26 | 27 |
| | 1996 | 18 | 26 | 26 | 29 |
| | 1997 | 13 | 24 | 36 | 27 |
| | 1998 | 9 | 25 | 39 | 27 |
| | 1999 | 15 | 24 | 35 | 26 |
| 11 | 1992 | 17 | 26 | 26 | 30 |
| | 1993 | 15 | 28 | 29 | 28 |
| | 1994 | 15 | 30 | 27 | 27 |
| | 1995 | 15 | 29 | 28 | 28 |
| | 1996 | 15 | 29 | 28 | 28 |
| | 1997 | 22 | 30 | 26 | 21 |
| | 1998 | 21 | 30 | 26 | 23 |
| | 1999 | 24 | 30 | 28 | 18 |

Note: Because of rounding, the percentages in each row may not total exactly to 100%.

**Table 9.5 (continued)**

SOCIAL SCIENCES

| Grade | Year | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|---|
| 4 | 1993 | 15 | 23 | 25 | 37 |
| | 1994 | 17 | 26 | 23 | 35 |
| | 1995 | 16 | 25 | 23 | 36 |
| | 1996 | 14 | 29 | 27 | 30 |
| | 1997 | 11 | 21 | 20 | 48 |
| | 1998 | 10 | 20 | 19 | 51 |
| | 1999 | 14 | 22 | 20 | 44 |
| 7 | 1993 | 19 | 22 | 22 | 37 |
| | 1994 | 22 | 22 | 20 | 36 |
| | 1995 | 21 | 24 | 22 | 33 |
| | 1996 | 20 | 24 | 22 | 34 |
| | 1997 | 16 | 16 | 18 | 49 |
| | 1998 | 15 | 17 | 16 | 52 |
| | 1999 | 19 | 20 | 15 | 47 |
| 11 | 1993 | 20 | 25 | 25 | 30 |
| | 1994 | 26 | 22 | 22 | 30 |
| | 1995 | 25 | 22 | 24 | 28 |
| | 1996 | 24 | 26 | 22 | 28 |
| | 1997 | 20 | 20 | 17 | 43 |
| | 1998 | 19 | 18 | 20 | 43 |
| | 1999 | 23 | 20 | 16 | 41 |

Note: Because of rounding, the percentages in each row may not total exactly to 100%.

# References

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement (3rd Edition)* (pp. 105-146). New York: Macmillan.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Peng, C-Y, J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359-368.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-non-mastery classifications. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 267-291). Baltimore: Johns Hopkins Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: Mesa.

# APPENDIX A. SUPPLEMENTARY TABLES

Tables A.1 through A.5 present correlations among the various standards sets, goal, or feature scores presented in student, school, and district reports. The sample sizes for the various analyses are summarized below. For writing at grades 5, 8, and 10, the sample size refers to the number of papers, not the number of students.

| | |
|---|---|
| Reading: Grade 3 | 142,743 |
| Reading: Grade 5 | 137,700 |
| Reading: Grade 8 | 134,672 |
| Reading: Grade 10 | 115,345 |
| | |
| Mathematics: Grade 3 | 144,288 |
| Mathematics: Grade 5 | 138,000 |
| Mathematics: Grade 8 | 134,458 |
| Mathematics: Grade 10 | 115,163 |
| | |
| Writing: Persuasive Prompt: Grade 3 | 45,764 |
| Writing: Expository Prompt: Grade 3 | 49,209 |
| Writing: Narrative Prompt: Grade 3 | 45,030 |
| | |
| Writing: Persuasive Prompt: Grade 5 | 136,749 |
| Writing: Expository Prompt: Grade 5 | 119,107 |
| Writing: Narrative Prompt: Grade 5 | 17,263 |
| | |
| Writing: Persuasive Prompt: Grade 8 | 134,553 |
| Writing: Expository Prompt: Grade 8 | 111,821 |
| Writing: Narrative Prompt: Grade 8 | 22,551 |
| | |
| Writing: Persuasive Prompt: Grade 10 | 116,421 |
| Writing: Expository Prompt: Grade 10 | 74,245 |
| Writing: Narrative Prompt: Grade 10 | 41,555 |
| | |
| Science: Grade 4 | 141,976 |
| Science: Grade 7 | 136,467 |
| Science: Grade 11 | 101,542 |
| | |
| Social Sciences: Grade 4 | 141,976 |
| Social Sciences: Grade 7 | 136,467 |
| Social Sciences: Grade 11 | 101,542 |

**Table A.1**
**Intercorrelations among Reading Standards Sets**

| Grade 3 | S1 | S2 | S3 | S4 | S5 | S6 |
|---|---|---|---|---|---|---|
| S1 | 1.000 | .698 | .829 | .966 | .753 | .581 |
| S2 | .698 | 1.000 | .845 | .813 | .706 | .571 |
| S3 | .829 | .845 | 1.000 | .800 | .704 | .557 |
| S4 | .966 | .813 | .800 | 1.000 | .781 | .610 |
| S5 | .753 | .706 | .704 | .781 | 1.000 | .610 |
| S6 | .581 | .571 | .557 | .610 | .610 | 1.000 |

| Grade 5 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| S1 | 1.000 | .744 | .798 | .886 | .700 |
| S2 | .744 | 1.000 | .888 | .938 | .740 |
| S3 | .798 | .888 | 1.000 | .807 | .695 |
| S4 | .886 | .938 | .807 | 1.000 | .761 |
| S5 | .700 | .740 | .695 | .761 | 1.000 |

| Grade 8 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| S1 | 1.000 | .670 | .813 | .942 | .764 |
| S2 | .670 | 1.000 | .837 | .813 | .682 |
| S3 | .813 | .837 | 1.000 | .764 | .701 |
| S4 | .942 | .813 | .764 | 1.000 | .781 |
| S5 | .764 | .682 | .701 | .781 | 1.000 |

| Grade 10 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| S1 | 1.000 | .693 | .756 | .911 | .725 |
| S2 | .693 | 1.000 | .815 | .882 | .728 |
| S3 | .756 | .815 | 1.000 | .755 | .688 |
| S4 | .911 | .882 | .755 | 1.000 | .782 |
| S5 | .725 | .728 | .688 | .782 | 1.000 |

**Table A.2**
**Intercorrelations among Mathematics Standards Sets**

| Grade 3 | S1 | S2 | | S4 | S5 | S6 | S7 | S8 |
|---|---|---|---|---|---|---|---|---|
| S1 | 1.000 | .938 | | .569 | .601 | .881 | .820 | .698 |
| S2 | .938 | 1.000 | | .518 | .554 | .788 | .741 | .643 |
| S4 | .569 | .518 | | 1.000 | .568 | .590 | .503 | .512 |
| S5 | .601 | .554 | | .568 | 1.000 | .604 | .543 | .532 |
| S6 | .881 | .788 | | .590 | .604 | 1.000 | .751 | .689 |
| S7 | .820 | .741 | | .503 | .543 | .751 | 1.000 | .629 |
| S8 | .698 | .643 | | .512 | .532 | .689 | .629 | 1.000 |

| Grade 5 | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 |
|---|---|---|---|---|---|---|---|---|
| S1 | 1.000 | .908 | .738 | .704 | .729 | .868 | .834 | .746 |
| S2 | .908 | 1.000 | .742 | .704 | .722 | .745 | .762 | .729 |
| S3 | .738 | .742 | 1.000 | .570 | .590 | .649 | .666 | .612 |
| S4 | .704 | .704 | .570 | 1.000 | .609 | .637 | .604 | .617 |
| S5 | .729 | .722 | .590 | .609 | 1.000 | .675 | .621 | .624 |
| S6 | .868 | .745 | .649 | .637 | .675 | 1.000 | .668 | .664 |
| S7 | .834 | .762 | .666 | .604 | .621 | .668 | 1.000 | .657 |
| S8 | .746 | .729 | .612 | .617 | .624 | .664 | .657 | 1.000 |

| Grade 8 | S1 | S2 | S3 | | S5 | S6 | S7 | S8 |
|---|---|---|---|---|---|---|---|---|
| S1 | 1.000 | .958 | .720 | | .841 | .867 | .825 | .737 |
| S2 | .958 | 1.000 | .704 | | .796 | .817 | .792 | .718 |
| S3 | .720 | .704 | 1.000 | | .656 | .653 | .642 | .591 |
| S5 | .841 | .796 | .656 | | 1.000 | .777 | .676 | .624 |
| S6 | .867 | .817 | .653 | | .777 | 1.000 | .675 | .625 |
| S7 | .825 | .792 | .642 | | .676 | .675 | 1.000 | .742 |
| S8 | .737 | .718 | .591 | | .624 | .625 | .742 | 1.000 |

| Grade 10 | S1 | S2 | S3 | | S5 | S6 | S7 | S8 |
|---|---|---|---|---|---|---|---|---|
| S1 | 1.000 | .965 | .790 | | .946 | .787 | .819 | .684 |
| S2 | .965 | 1.000 | .823 | | .941 | .806 | .821 | .700 |
| S3 | .790 | .823 | 1.000 | | .772 | .716 | .710 | .593 |
| S5 | .946 | .941 | .772 | | 1.000 | .769 | .745 | .639 |
| S6 | .787 | .806 | .716 | | .769 | 1.000 | .691 | .589 |
| S7 | .819 | .821 | .710 | | .745 | .691 | 1.000 | .708 |
| S8 | .684 | .700 | .593 | | .639 | .589 | .708 | 1.000 |

**Table A.3**
**Intercorrelations among Writing Feature Scores**

| Persuasive Prompt: Grade 3 | F | S | O | C | I |
|---|---|---|---|---|---|
| F | 1.000 | .743 | .767 | .485 | .866 |
| S | 743 | 1.000 | .894 | .420 | .907 |
| O | 767 | .894 | 1.000 | .437 | .921 |
| C | 485 | .420 | .437 | 1.000 | .463 |
| I | 866 | .907 | .921 | .463 | 1.000 |
| | | | | | |
| Expository Prompt: Grade 3 | F | S | O | C | I |
| F | 1.000 | .729 | .762 | .462 | .861 |
| S | 729 | 1.000 | .886 | .410 | .902 |
| O | 762 | .886 | 1.000 | .422 | .916 |
| C | 462 | .410 | .422 | 1.000 | .448 |
| I | 861 | .902 | .916 | .448 | 1.000 |
| | | | | | |
| Narrative Prompt: Grade 3 | F | S | O | C | I |
| F | 1.000 | .817 | .841 | .262 | .906 |
| S | 817 | 1.000 | .846 | .257 | .912 |
| O | 841 | .846 | 1.000 | .281 | .927 |
| C | 262 | .257 | .281 | 1.000 | .284 |
| I | 906 | .912 | .927 | .284 | 1.000 |
| | | | | | |
| Persuasive Prompt: Grade 5 | F | S | O | C | I |
| F | 1.000 | .727 | .792 | .293 | .846 |
| S | .727 | 1.000 | .826 | .283 | .901 |
| O | .792 | .826 | 1.000 | .304 | .931 |
| C | .293 | .283 | .304 | 1.000 | .311 |
| I | .846 | .901 | .931 | .311 | 1.000 |
| | | | | | |
| Expository Prompt: Grade 5 | F | S | O | C | I |
| F | 1.000 | .691 | .768 | .282 | .825 |
| S | .691 | 1.000 | .808 | .266 | .890 |
| O | .768 | .808 | 1.000 | .304 | .925 |
| C | .282 | .266 | .304 | 1.000 | .302 |
| I | .825 | .890 | .925 | .302 | 1.000 |

**Table A.3 (continued)**

| Narrative Prompt: Grade 5 | F | S | O | C | I |
|---|---|---|---|---|---|
| F | 1.000 | .785 | .857 | .225 | .920 |
| S | .785 | 1.000 | .796 | .233 | .878 |
| O | .857 | .796 | 1.000 | .248 | .926 |
| C | .225 | .233 | .248 | 1.000 | .242 |
| I | .920 | .878 | .926 | .242 | 1.000 |
| **Persuasive Prompt: Grade 8** | **F** | **S** | **O** | **C** | **I** |
| F | 1.000 | .685 | .727 | .365 | .781 |
| S | .685 | 1.000 | .843 | .378 | .898 |
| O | .727 | .843 | 1.000 | .416 | .935 |
| C | .365 | .378 | .416 | 1.000 | .417 |
| I | .781 | .898 | .935 | .417 | 1.000 |
| **Expository Prompt: Grade 8** | **F** | **S** | **O** | **C** | **I** |
| F | 1.000 | .665 | .714 | .337 | .768 |
| S | .665 | 1.000 | .827 | .334 | .889 |
| O | .714 | .827 | 1.000 | .381 | .929 |
| C | .337 | .334 | .381 | 1.000 | .379 |
| I | .768 | .889 | .929 | .379 | 1.000 |
| **Narrative Prompt: Grade 8** | **F** | **S** | **O** | **C** | **I** |
| F | 1.000 | .808 | .897 | .344 | .942 |
| S | .808 | 1.000 | .808 | .360 | .878 |
| O | .897 | .808 | 1.000 | .360 | .937 |
| C | .344 | .360 | .360 | 1.000 | .368 |
| I | .942 | .878 | .937 | .368 | 1.000 |
| **Persuasive Prompt: Grade 10** | **F** | **S** | **O** | **C** | **I** |
| F | 1.000 | .787 | .803 | .391 | .833 |
| S | .787 | 1.000 | .941 | .381 | .960 |
| O | .803 | .941 | 1.000 | .390 | .972 |
| C | .391 | .381 | .390 | 1.000 | .394 |
| I | .833 | .960 | .972 | .394 | 1.000 |

**Table A.3 (continued)**

| Expository Prompt: Grade 10 | F | S | O | C | I |
|---|---|---|---|---|---|
| F | 1.000 | .780 | .799 | .386 | .825 |
| S | .780 | 1.000 | .938 | .372 | .959 |
| O | .799 | .938 | 1.000 | .384 | .971 |
| C | .386 | .372 | .384 | 1.000 | .386 |
| I | .825 | .959 | .971 | .386 | 1.000 |
| **Narrative Prompt: Grade 10** | **F** | **S** | **O** | **C** | **I** |
| F | 1.000 | .861 | .909 | .275 | .948 |
| S | .861 | 1.000 | .871 | .291 | .914 |
| O | .909 | .871 | 1.000 | .297 | .956 |
| C | .275 | .291 | .297 | 1.000 | .293 |
| I | .948 | .914 | .956 | .293 | 1.000 |

**Table A.4**
**Intercorrelations among Science Goal Scores**

| Grade 4 | G1 | G2 | G3 | G4 |
|---|---|---|---|---|
| G1 | 1.000 | .744 | .724 | .697 |
| G2 | .744 | 1.000 | .746 | .689 |
| G3 | .724 | .746 | 1.000 | .705 |
| G4 | .697 | .689 | .705 | 1.000 |
| **Grade 7** | **G1** | **G2** | **G3** | **G4** |
| G1 | 1.000 | .695 | .653 | .689 |
| G2 | .695 | 1.000 | .721 | .721 |
| G3 | .653 | .721 | 1.000 | .692 |
| G4 | .689 | .721 | .692 | 1.000 |
| **Grade 11** | **G1** | **G2** | **G3** | **G4** |
| G1 | 1.000 | .709 | .696 | .732 |
| G2 | .709 | 1.000 | .671 | .669 |
| G3 | .696 | .671 | 1.000 | .678 |
| G4 | .732 | .669 | .678 | 1.000 |

**Table A.5**
**Intercorrelations among Social Sciences Goal Scores**

| Grade 4 | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| G1 | 1.000 | .725 | .659 | .681 | .725 |
| G2 | .725 | 1.000 | .676 | .680 | .720 |
| G3 | .659 | .676 | 1.000 | .613 | .652 |
| G4 | .681 | .680 | .613 | 1.000 | .698 |
| G5 | .725 | .720 | .652 | .698 | 1.000 |
| | | | | | |
| Grade 7 | G1 | G2 | G3 | G4 | G5 |
| | | | | | |
| G1 | 1.000 | .739 | .634 | .680 | .706 |
| G2 | .739 | 1.000 | .613 | .687 | .694 |
| G3 | .634 | .613 | 1.000 | .570 | .615 |
| G4 | .680 | .687 | .570 | 1.000 | .672 |
| G5 | .706 | .694 | .615 | .672 | 1.000 |
| | | | | | |
| Grade 11 | G1 | G2 | G3 | G4 | G5 |
| | | | | | |
| G1 | 1.000 | .694 | .597 | .652 | .675 |
| G2 | .694 | 1.000 | .586 | .661 | .680 |
| G3 | .597 | .586 | 1.000 | .558 | .583 |
| G4 | .652 | .661 | .558 | 1.000 | .645 |
| G5 | .675 | .680 | .583 | .645 | 1.000 |