The Illinois State Assessment
2001 Technical Manual

Illinois State Board of Education
Division of Assessment

CONTENTS

1. PURPOSE AND DESIGN OF THE ISAT TESTING PROGRAM

In April 2001, students in grades 3, 5, and 8 took Illinois Standards Achievement Tests (ISAT) in reading, mathematics, and writing. Students in grades 4 and 7 took ISAT tests in science and social science. Approximately 700,000 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. 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 2001 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 2001 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 40-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 extended-response question.

The reading tests for grades 5 and 8 are also given in three 40-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 extended-response 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 extended-response 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 extended-response items 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.

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 nonfiction 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 70 multiple-choice items administered in two test sessions. A third session contains two extended-response/problem-solving tasks.

In addition to an overall mathematics score, results are reported in terms of the percent of items correctly answered within eight standard sets. 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, 8D)

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

- *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. (Standard 9A)

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

- *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 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 and 8 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/elaboration, 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/Elaboration + 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.

Science

Science is a creative endeavor of the human mind. It offers a special perspective on the natural world in terms of understanding and interaction. The Illinois Learning Standards for science are organized by goals that inform one another and depend upon one another for meaning. Expectations for learners related to the inquiry process are presented in standards addressing the doing of science and elements of technological design.

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

- Goal 11: Understand the process of scientific inquiry and technological design to investigate questions, conduct experiments, and solve problems.
  11A. Know and apply the concepts, principles and processes of scientific inquiry.
  11B. Know and apply the concepts, principles, and processes of technological design.

- Goal 12: Understand the fundamental concepts, principles, and interconnections of the life, physical, and earth/space sciences.
  12A. Know and apply concepts that explain how living things function, adapt, and change.
  12B. Know and apply concepts that describe how living things interact with each other and with their environment.
  12C. Know and apply concepts that describe properties of matter and energy and the interactions between them.
  12D. Know and apply concepts that describe force and motion and the principles that explain them.
  12E. Know and apply concepts that describe the features and processes of the earth and its resources.

12F. Know and apply concepts that explain the composition and structure of the universe and earth's place in it.

- Goal 13: Understand the relationships among science, technology, and society in historical and contemporary contexts.
  13A. Know and apply the accepted practices of science.
  13B. Know and apply concepts that describe the interaction between science, technology, and society.

The science assessment consists of single-correct-answer, multiple-choice items. In addition to an overall score, results are reported in terms of the percent of items correctly answered within five standard sets. These scores are as follows:

- *Scientific Inquiry:* Understanding and applying knowledge of experimental and technological design, including data analysis, use of scientific instruments, and the metric system. (Standards 11A and 11B)

- *Life Sciences:* Understanding and applying knowledge of biology and ecology. (Standards 12A and 12B)

- *Physical Sciences:* Understanding and applying knowledge of chemistry and physics. (Standards 12C and 12D)

- *Earth and Space Sciences:* Understanding and applying knowledge of geology, weather, renewable resources, astronomy, and space science. (Standards 12E and 12F)

- *Science, Technology, and Society:* Understanding and applying knowledge of safety, valid sources of data, and ethical practices. Understanding and applying knowledge of the history and sociology of science, ethics, environmental issues, and recycling. (Standards 13A and 13B)

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 of items within the tests among major learning areas: earth science, physical science, and life science.

Social Science

Social science provides students with an understanding of themselves and of society, prepares them for citizenship in a democracy, and gives them the basics for understanding the complexities of the world community. The study of social science helps people develop the ability to make informed and reasoned decisions for the public good as citizens of a culturally diverse, democratic society in an interdependent world.

The ISAT social science tests are designed to measure the following learning standards:

- Goal 14: Understand political systems with an emphasis on the United States.
  14A. Understand and explain basic principles of the United States government.
  14B. Understand the structures and functions of the political systems of Illinois, the United States, and other nations.
  14C. Understand election processes and responsibilities of citizens.
  14D. Understand the roles and influences of individuals and interest groups in the political systems of Illinois, the United States, and other nations.
  14E. Understand United States foreign policy as it relates to other nations and international issues.
  14F. Understand the development of United States political ideas and traditions.

- Goal 15: Understand economic systems with an emphasis on the United States.
  15A. Understand how different economic systems operate in the exchange, production, distribution, and consumption of goods and services.
  15B. Understand that scarcity necessitates choices by consumers.
  15C. Understand that scarcity necessitates choices by producers.
  15D. Understand trade as an exchange of goods or services.
  15E. Understand the impact of government policies and decisions on production and consumption in the economy.

- Goal 16: Understand events, trends, individuals, and movements shaping the history of Illinois, the United States, and other nations.
  16A. Apply the skills of historical analysis and interpretation.
  16B. Understand the development of significant political events.
  16C. Understand the development of economic systems.
  16D. Understand Illinois, United States, and world social history.
  16E. Understand Illinois, United States, and world environmental history.

- Goal 17: Understand world geography and the effects of geography on society, with an emphasis on the United States.
  17A. Locate, describe, and explain places, regions, and features on the Earth.
  17B. Analyze and explain characteristics and interactions of the Earth's physical systems.
  17C. Understand relationships between geographic factors and society.
  17D. Understand the historical significance of geography.

- Goal 18: Understand social systems with an emphasis on the United States.
  18A. Compare characteristics of culture as reflected in language, literature, the arts, traditions, and institutions.
  18B. Understand the roles and interactions of individuals and groups in society.
  18C. Understand how social systems form and develop over time.

The social science assessment consists of single-correct-answer, multiple-choice items. In addition to an overall score, results are reported in terms of the percent of items correctly answered within five standard sets. These scores are as follows:

- *Government:* Understanding and applying knowledge of political systems, including the basic principles and traditions of the U.S. government, the structure and functions of government, the election process, and foreign policy. (Standards 14A, 14B, 14D, 14F, and 18B)

- *Economics:* Understanding and applying knowledge of economic systems and the nature of the U.S. economy, including the choices people make in the production and distribution of goods and services and the relationship of governments to trade and economic practices. (Standards 15A, 15B, 15C, 15D, and 15E)

- *Geography:* Demonstrating the ability to locate places, regions, and features; to understand characteristics of Earth's physical system and the relationship between geographic factors and society; and to understand the historical significance of geography. (Standards 17A, 17B, 17C, and 17D)

- *United States History:* Understanding and analyzing the development of political events, economic systems, and social systems. (Standards 16A, 16B, 16C, 16D, 16E, 18A, 18B, and 18C)

- *Global Perspectives:* Understanding and applying knowledge of the political, economic, historical, social, and environmental events and conditions in the world beyond the United States. (Standards 14B, 14E, 16A, 16B, 16C, 16D, 16E, 18A, 18B, and 18C)

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 2001 assessment.

Table 2.1
2001 Reliability Estimates

Grade	Reading	Mathematics	Writing	Science	Social Science
03	.94	.94	.87		
04				.93	.93
05	.94	.94	.90		
07				.92	.91
08	.93	.95	.91		

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

Reading: 3 (15,775), 5 (15,175), 8 (15,250)
Mathematics: 3 (15,873), 5 (15,829), 8 (15,723)
Writing: 3 (140,574), 5 (150,958), 8 (137,345)
Science: 4 (15,888), 7 (15,835)
Social Science: 4 (15,902), 7 (15,856)

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; Figures 2.2, 2.3, and 2.4 present comparable information for the ISAT mathematics tests, science tests, and social science tests, respectively. 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. As is evident from the figures, the information functions for these tests are highest near the points on the scales where the "meets standards" cut scores are located.

Figure 2.1
2001 ISAT Reading Test Information Functions

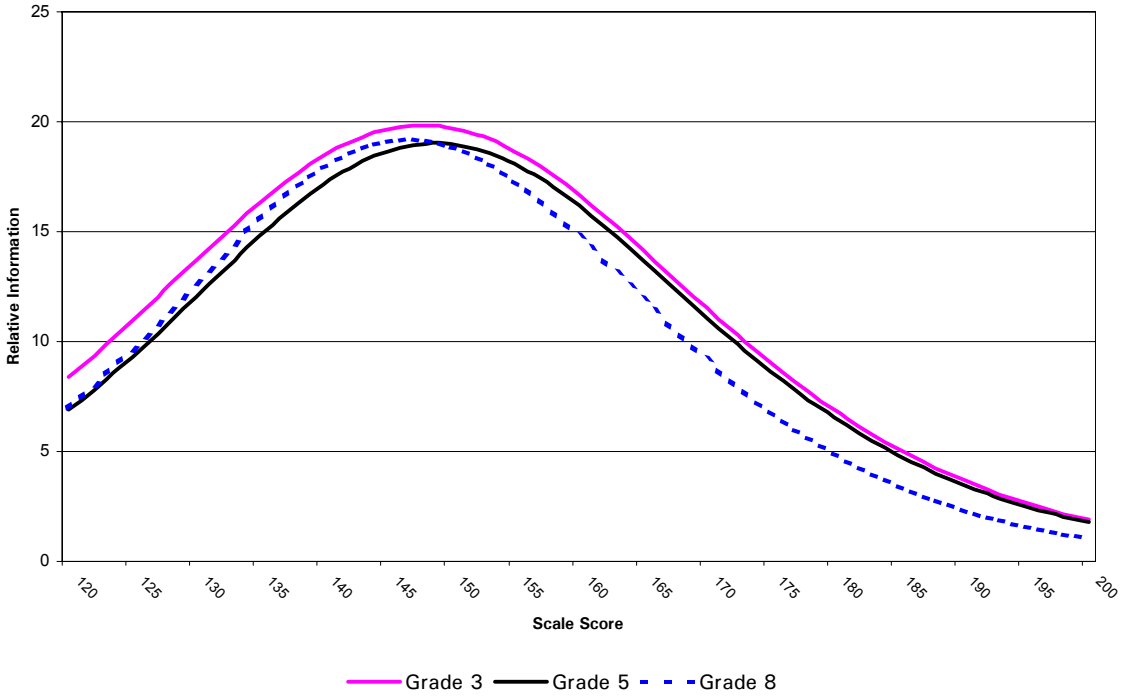


Figure 2.2
2001 ISAT Mathematics Test Information Functions

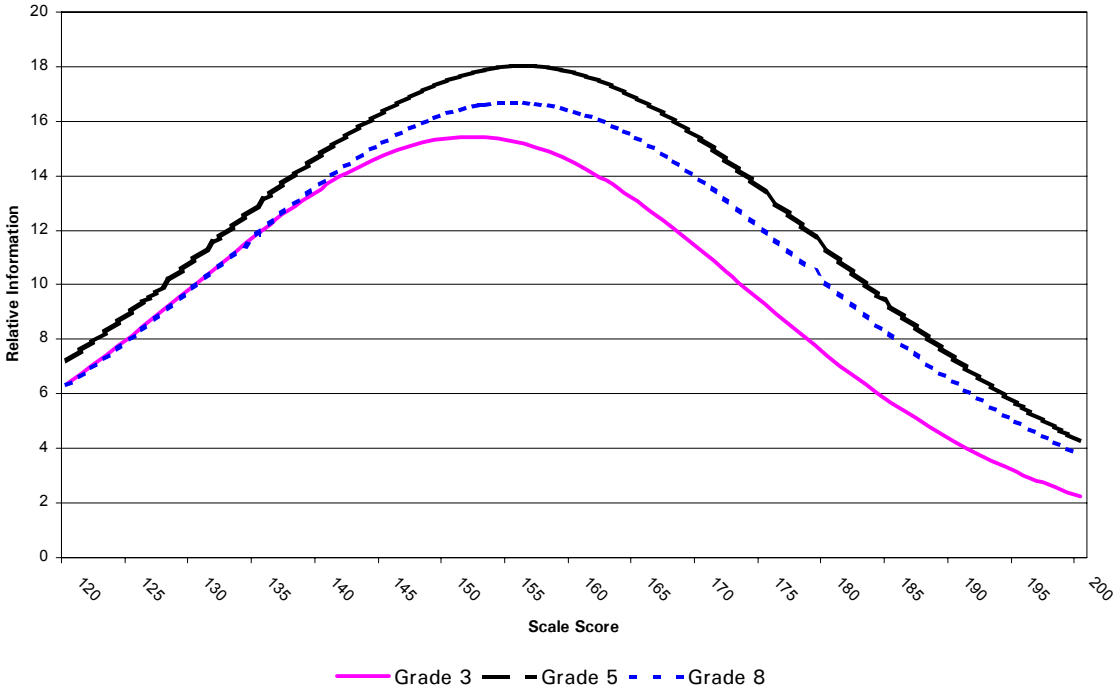


Figure 2.3
2001 ISAT Science Test Information Functions

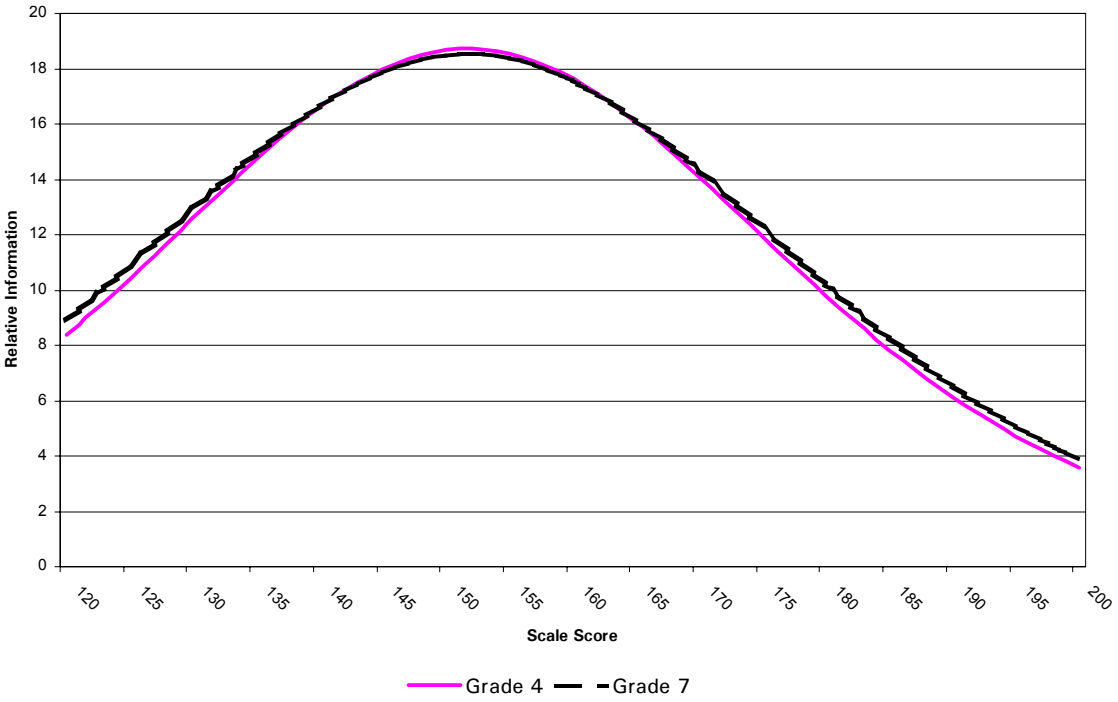
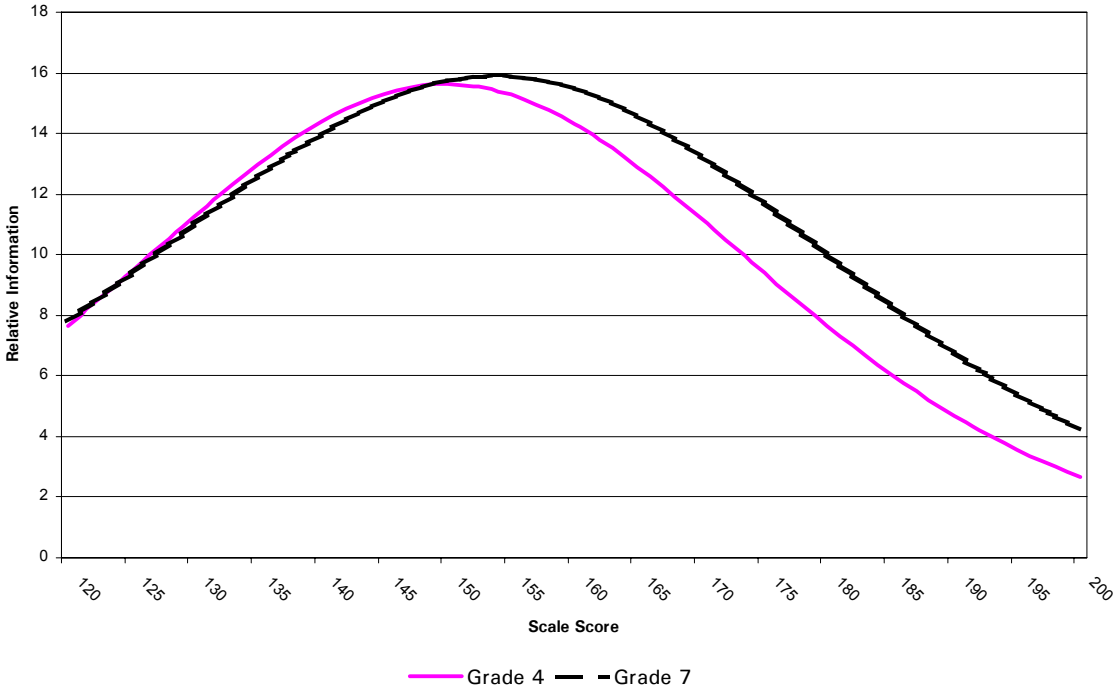


Figure 2.4
2001 ISAT Social Science Test Information Functions



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. Person separation values of 3 and above indicate a high degree of measurement precision. As the table indicates, the ISAT reading, mathematics, science, and social science tests show consistently high levels of test precision across all the grade levels tested. Person separation values for the reading and mathematics tests are exceptionally high.

Table 2.2
Person Separation Values for the 2001 ISAT Tests

	Reading	Mathematics
Grade 3	3.42	3.66
Grade 5	3.59	3.36
Grade 8	3.21	3.92
	Science	Social Science
Grade 4	3.32	3.31
Grade 7	3.18	3.09

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- and 8th-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 2001 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 94% of the time.

Table 2.3
Interrater Agreement for Writing Scores

Discourse Mode	Score	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
Persuasive	Focus	63	33	96
	Support	60	37	97
(n = 23,784)	Organization	61	37	98
	Conventions	95	5	100
	Integration	62	36	98
Expository	Focus	63	31	94
	Support	57	39	96
(n = 5,723)	Organization	60	37	97
	Conventions	95	5	100
	Integration	59	38	97
Narrative	Focus	61	36	97
	Support	61	37	98
(n = 28,595)	Organization	61	36	97
	Conventions	94	6	100
	Integration	63	35	98

A set of validation papers was also developed for monitoring scoring of extended responses in reading and mathematics. For the reading test, raters provided a single score for the extended-response item, while extended-response items in the mathematics test were scored for knowledge, strategy, and explanation. Tables 2.4 and 2.5 present interrater agreement statistics for extended responses in reading and mathematics, respectively.

Table 2.4
Interrater Agreement for Reading Extended-Response Items

	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
Grade 3			
(N = 12,920)			
Item 1	66	31	97
Item 2	60	37	97
Grade 5			
(N = 13,820)			
Item 1	65	34	99
Item 2	71	28	99
Grade 8			
(N = 12,368)			
Item 1	75	25	100
Item 2	75	25	100

Table 2.5
Interrater Agreement for Mathematics Extended-Response Items

	Score	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
Grade 3				
Task 1	Knowledge	65	29	94
(N = 13,613)	Strategy	63	29	92
	Explanation	57	31	88
Task 2	Knowledge	87	11	98
	Strategy	82	15	97
	Explanation	60	31	91
Grade 5				
Task 1	Knowledge	94	4	98
(N = 14,441)	Strategy	93	5	98
	Explanation	62	31	93
Task 2	Knowledge	79	19	98
	Strategy	79	15	94
	Explanation	65	30	95
Grade 8				
Task 1	Knowledge	82	18	100
(N = 13,451)	Strategy	77	19	96
	Explanation	61	32	93
Task 2	Knowledge	73	24	97
	Strategy	68	26	94
	Explanation	64	30	94

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.

Table 2.6
Agreement with Validation Papers for Writing Scores

Discourse Mode	Score	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
Grade 3				
Narrative	Focus	80	19	99
	Support	83	17	100
	Organization	81	19	100
(N = 470)	Conventions	91	9	100
	Integration	82	17	99
Persuasive	Focus	63	29	92
/Expository	Support	59	38	97
(N = 1,980)	Organization	61	36	97
	Conventions	89	11	100
	Integration	59	38	97
Grade 5				
Narrative	Focus	65	31	96
	Support	61	37	98
	Organization	63	34	97
(N = 2,020)	Conventions	97	4	101
	Integration	61	36	97
Persuasive	Focus	75	20	95
/Expository	Support	68	31	99
(N = 1,960)	Organization	68	30	98
	Conventions	94	6	100
	Integration	67	31	98
Grade 8				
Narrative	Focus	76	24	100
	Support	80	20	100
	Organization	81	19	100
(N = 1,630)	Conventions	93	7	100
	Integration	81	19	100
Persuasive	Focus	70	28	98
/Expository	Support	62	36	98
(N = 2,350)	Organization	64	34	98
	Conventions	96	4	100
	Integration	64	34	98

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 grades, features, and discourse modes are summarized in Table 2.6. 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.

A set of validation papers was also developed for monitoring scoring of extended responses in reading and mathematics. For the reading test, raters provided a single score for the extended-response item, while extended-response items in the mathematics test were scored for knowledge, strategy, and explanation. Tables 2.7 and 2.8 present agreement with validation papers for extended responses in reading and mathematics, respectively.

Table 2.7
Agreement with Validation Papers for Reading Extended-Response Items

	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
Grade 3			
Item 1 (N = 1,240)	70	29	99
Item 2 (N = 1,240)	68	29	97
Grade 5			
Item 1 (N = 1,984)	77	23	100
Item 2 (N = 1,984)	77	23	100
Grade 8			
Item 1 (N = 1,728)	84	15	99
Item 2 (N = 1,728)	79	20	99

Table 2.8
Agreement with Validation Papers for Mathematics Extended-Response Items

	Score	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
Grade 3				
Task 1	Knowledge	84	13	97
(N = 735)	Strategy	84	12	96
	Explanation	78	13	91
Task 2	Knowledge	96	4	100
(N = 735)	Strategy	95	4	99
	Explanation	79	15	94
Grade 5				
Task 1	Knowledge	97	1	98
(N = 1,240)	Strategy	97	2	99
	Explanation	74	23	97
Task 2	Knowledge	93	6	99
(N = 1,240)	Strategy	93	6	99
	Explanation	84	13	97
Grade 8				
Task 1	Knowledge	88	12	100
(N = 1,200)	Strategy	89	9	98
	Explanation	67	25	92
Task 2	Knowledge	80	20	100
(N = 1,200)	Strategy	73	24	97
	Explanation	71	24	95

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 were developed for each learning area and each grade. The development of the score cutoffs that define these categories is fully documented in separate publications available from ISBE (*Performance Levels for the Illinois Standards Achievement Tests: Reading, Mathematics, Writing* and *Performance Levels for the Illinois Standards Achievement Tests: Science, Social Science*). 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 (reading, mathematics, writing) and April 2000 (science, social science), 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; science—30; social science—30.

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, mathematics, science, and social science 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 2.9. Results of applying these cutoffs to the 2001 test population are shown later in Section 4.

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.

Table 2.7
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
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
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
SCIENCE	**Academic Warning**	**Below Standards**	**Meets Standards**	**Exceeds Standards**
04	120-138	139-153	154-178	179-200
07	120-141	142-150	151-174	175-200
SOCIAL SCIENCE	**Academic Warning**	**Below Standards**	**Meets Standards**	**Exceeds Standards**
04	120-141	142-156	157-183	184-200
07	120-132	133-156	157-178	179-200

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 2001 tests.

Table 2.10 presents the 2001 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 2.10
Reliability of Student Performance Decisions Based on 2001 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	.964	.737	.036	.902	.796	.098	.936	.774	.064
	5	.990	.757	.010	.906	.797	.094	.918	.786	.082
	8	.992	.746	.008	.910	.795	.090	.948	.762	.052
Mathematics	3	.964	.737	.036	.924	.783	.076	.916	.789	.084
	5	.978	.714	.022	.936	.774	.064	.918	.786	.082
	8	.968	.744	.032	.926	.779	.074	.922	.787	.078
Writing	3	.926	.578	.074	.824	.648	.176	.922	.588	.078
	5	.986	.604	.014	.878	.704	.122	.918	.673	.082
	8	.966	.604	.034	.866	.710	.134	.930	.656	.070
Science	4	.964	.737	.036	.910	.795	.090	.950	.754	.050
	7	.942	.642	.058	.892	.692	.108	.902	.691	.098
Social Science	4	.954	.757	.046	.902	.796	.098	.970	.738	.030
	7	.982	.599	.018	.866	.704	.134	.918	.673	.082
AVERAGE		.961	.684	.040	.893	.747	.108	.935	.714	.065

3. SCALING AND EQUATING PROCEDURES

ISAT reading, mathematics, science, and social science 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 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 are 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 Scaling Constants

	Slope	Intercept
Reading		
Grade 3	12.6428	146.2066
Grade 5	12.0100	144.7660
Grade 8	11.2280	141.7730
Mathematics		
Grade 3	13.5122	147.6910
Grade 5	14.9686	153.4644
Grade 8	14.7578	146.7806
Science		
Grade 4	15.3781	152.4255
Grade 7	15.9209	152.4527
Social Science		
Grade 4	14.6746	149.2394
Grade 7	16.6587	148.9095

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. 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.

The statistical fit of the one-parameter logistic (1PL) or Rasch model to the ISAT multiple-choice tests has been previously examined and found to be satisfactory. The 1PL model uses only the item difficulty and the person's proficiency level to describe the probability of a correct response to an item. The 1PL model is the simplest of currently available IRT models and is perhaps the one in widest use today.

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 ISAT 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 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 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.

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 also uses two forms of the reading test. 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.

Tables 3.2 through 3.4 show results of the Rasch calibration and equating procedures for reading. Column 1 of each table shows the Form in which the item appeared, A for Form A alone, B for Form B alone, AB for items that appeared in both forms. Column 2 of each table shows the item number within the test booklet. Column 3 shows the Rasch difficulties resulting from an anchored (constrained) calibration of the 2001 test. Column 4 shows the standard error of the difficulty estimate (S_{ed}). The next two columns present statistics designed to assess how well the test "fits" the IRT model. Both are standardized, mean-square statistics with an expected value of 1.00 (indicating perfect fit). The first, "Infit," is more sensitive to departures from model fit when item difficulty and person ability are close. The second, "Outfit," is more sensitive to model fit when item difficulty and person

ability are far apart. The last column shows the point-biserial correlation between the item and the rest of the items in the test.

Tables 3.5 through 3.7 show similar information for the mathematics tests. The information is organized in the same way as the earlier tables except no "form" designation is necessary. Tables 3.8 and 3.9 present information for the science tests, and Tables 3.10 and 3.11 present information for the social science tests.

Table 3.2
Results of the 2001 Equating Process–Reading Grade 3

Form	Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
AB	1	-1.27	.03	.95	.84	.37
AB	2	-.19	.02	1.22	1.28	.25
AB	3	.53	.02	1.04	1.02	.43
AB	4	-.18	.02	1.08	1.12	.36
AB	5	-.99	.02	.99	.93	.36
AB	6	-.34	.02	1.04	1.02	.38
AB	7	-.59	.02	.97	.93	.41
AB	8	.20	.02	.90	.86	.52
AB	9	-.95	.02	.99	1.00	.36
AB	10	.60	.02	1.15	1.20	.34
AB	11	.92	.02	1.19	1.25	.32
AB	12	.73	.02	1.23	1.35	.28
AB	13	-.19	.02	1.20	1.22	.27
AB	14	.59	.02	1.06	1.04	.42
AB	15	.71	.02	1.19	1.26	.31
AB	16	-.73	.02	.90	.78	.46
AB	17	.93	.02	1.05	1.06	.43
AB	18	.91	.02	1.04	1.04	.44
AB	19	-.22	.02	.94	.90	.47
AB	20	.29	.02	1.01	1.10	.43
AB	21	-.52	.02	.92	.82	.46
AB	22	-1.13	.02	.91	.76	.41
AB	23	.33	.02	.97	.93	.47
AB	24	.94	.02	.98	.96	.48
AB	25	.09	.02	.98	.92	.46
AB	26	.57	.02	1.11	1.15	.37
AB	27	1.70	.02	1.20	1.42	.28
AB	28	.25	.02	1.02	1.02	.43
AB	29	-1.13	.02	.95	.93	.37
AB	30	-.52	.02	.90	.76	.48
AB	31	-.84	.02	.96	.95	.40
AB	32	-.21	.02	.81	.69	.57
AB	33	.15	.02	.94	.85	.50
AB	34	-.13	.02	.95	.87	.47
AB	35	-.07	.02	1.02	1.04	.41
AB	36	-.68	.02	.88	.76	.48
AB	37	-1.18	.03	.87	.65	.45
AB	38	-.28	.02	.92	.85	.48
AB	39	1.25	.02	1.07	1.13	.40
AB	40	.58	.02	1.02	1.01	.45
AB	41	.19	.02	.91	.86	.51
AB	42	.42	.02	.99	.94	.47
AB	43	.27	.02	.90	.82	.53
AB	44	.33	.02	.94	.86	.50
AB	45	-1.11	.02	.83	.55	.49
AB	46	.14	.02	.95	.95	.47
AB	47	.94	.02	1.09	1.12	.40
A*	48	.10	.03	.98	.90	.49
A*	49	-.50	.03	.97	.98	.42

Table 3.2 (continued)

Form	Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
A	50	.99	.03	.97	.98	.48
A	51	-.81	.03	.89	.81	.46
A	52	1.17	.03	1.35	1.48	.19
A	53	-.74	.03	.81	.61	.52
A	54	-.41	.03	.91	.75	.50
A	55	.96	.03	1.12	1.16	.37
A	56	.44	.03	.87	.78	.54
A	57	.10	.03	.84	.74	.57
A	58	.15	.03	.93	.88	.49
A	59	.26	.03	1.02	1.12	.40
A	60	.39	.03	.84	.75	.56
A	61	.73	.03	1.08	1.10	.40
A	62	1.22	.03	1.16	1.23	.33
A	63	-.50	.03	.85	.71	.51
A	64	1.29	.03	1.04	1.10	.43
A	65	.12	.03	.90	.84	.53
A	66	.08	.03	.86	.80	.52
A	67	.47	.03	.99	.96	.44
B	68	-.14	.03	.94	.89	.47
B	69	-.21	.03	1.00	.98	.42
B	70	-.12	.03	.87	.73	.54
B	71	.18	.03	1.09	1.12	.38
B	72	.03	.03	.96	.92	.47
B	73	-.36	.03	.95	.88	.45
B	74	.27	.03	.95	.86	.50
B	75	-.16	.03	.94	.90	.47
B	76	-.10	.03	.95	.89	.47
B	77	-.31	.03	.89	.79	.50
B	78	.17	.03	.97	.94	.47
B	79	.60	.03	.95	.93	.50
B	80	.33	.03	.91	.87	.52
B	81	.20	.03	.99	.97	.45
B	82	-.21	.03	.92	.86	.48
B	83	1.07	.03	1.09	1.11	.40
B	84	-.10	.03	.84	.73	.55
B	85	-.02	.03	.95	.88	.47
B	86	-.66	.03	.90	.77	.47
B	87	.96	.03	1.17	1.21	.34
ER	88	2.89	.01	1.27	1.37	.45
ER	89	2.54	.02	1.17	1.16	.52
ER	90	1.99	.02	1.10	1.09	.58

Note: ER = Extended-response item.

Table 3.3
Results of the 2001 Equating Process–Reading Grade 5

Form	Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
A	1	-.11	.03	1.00	.95	.40
A	2	-.40	.03	1.25	1.61	.21
A	3	2.08	.03	1.19	1.37	.38
A	4	.99	.03	1.24	1.32	.31
A	5	.42	.03	.86	.77	.56
A	6	.39	.03	.95	.89	.48
A	7	-.34	.03	1.02	.99	.39
A	8	.79	.03	1.08	1.08	.42
A	9	.57	.03	.98	.90	.47
A	10	.04	.03	.99	.94	.44
A	11	1.09	.03	1.11	1.16	.40
A	12	.64	.03	1.12	1.21	.38
A	13	.08	.03	1.09	1.05	.39
A	14	1.26	.03	1.09	1.11	.42
A	15	.86	.03	.93	.87	.51
A	16	1.46	.03	1.04	1.08	.46
A	17	1.87	.03	1.16	1.30	.33
A	18	.63	.03	1.03	1.01	.45
AB	19	-.20	.02	.90	.84	.49
AB	20	-.32	.02	.94	.88	.46
AB	21	1.00	.02	1.18	1.25	.36
AB	22	-.34	.02	.89	.78	.49
AB	23	.31	.02	1.01	.98	.46
AB	24	-.44	.02	.88	.71	.50
AB	25	-.20	.02	.86	.73	.52
AB	26	.52	.02	.95	.88	.51
AB	27	.51	.02	1.02	.97	.46
AB	28	.34	.02	.87	.75	.56
AB	29	-.50	.02	.87	.75	.49
AB	30	.27	.02	1.02	1.06	.44
AB	31	.54	.02	.95	.89	.51
AB	32	.26	.02	.88	.81	.54
AB	33	.82	.02	.98	.95	.50
AB	34	.64	.02	1.04	1.02	.45
AB	35	.03	.02	.91	.84	.50
AB	36	1.03	.02	.98	.97	.50
AB	37	-.13	.02	.91	.80	.50
AB	38	.68	.02	.99	1.00	.48
AB	39	-.81	.02	.91	.81	.44
AB	40	-.35	.02	.92	.86	.47
AB	41	1.15	.02	1.08	1.11	.43
AB	42	1.03	.02	1.17	1.26	.37
AB	43	.32	.02	1.04	1.10	.42
AB	44	-.01	.02	.94	.93	.48
AB	45	.18	.02	.98	.96	.46
AB	46	1.20	.02	.96	.94	.52
AB	47	.36	.02	1.01	.99	.46
AB	48	1.41	.02	1.09	1.14	.43
AB	49	.76	.02	.91	.84	.55

Table 3.3 (continued)

Form	Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
AB	50	-.34	.02	.92	.84	.47
AB	51	-.25	.02	1.00	1.07	.41
AB	52	-.31	.02	.88	.81	.50
AB	53	.01	.02	.97	1.02	.45
AB	54	.31	.02	.89	.81	.54
AB	55	1.12	.02	1.06	1.10	.44
AB	56	-.33	.02	.98	.98	.42
AB	57	.44	.02	1.10	1.11	.40
AB	58	1.68	.02	1.22	1.37	.33
AB	59	1.55	.02	1.09	1.17	.42
AB	60	.33	.02	.93	.90	.51
AB	61	.67	.02	1.04	1.06	.45
AB	62	.62	.02	1.10	1.20	.40
AB	63	-.48	.02	.93	.84	.45
AB	64	-.07	.02	.84	.70	.55
AB	65	-.10	.02	.94	.89	.47
AB	66	.73	.02	1.37	1.68	.20
B	67	-1.37	.04	.92	.74	.38
B	68	-.11	.03	1.01	.96	.44
B	69	.72	.03	1.00	.97	.49
B	70	.06	.03	1.06	1.04	.42
B	71	-.37	.03	.97	.92	.44
B	72	.14	.03	1.03	1.04	.44
B	73	.16	.03	1.09	1.15	.39
B	74	.72	.03	.94	.89	.53
B	75	-.13	.03	.92	.82	.49
B	76	.01	.03	1.00	.96	.45
B	77	.20	.03	.97	.95	.48
B	78	.62	.03	1.02	1.01	.48
B	79	-.59	.03	.85	.69	.50
B	80	-1.03	.04	.94	.96	.39
B	81	.79	.03	1.17	1.24	.38
B	82	.59	.03	.94	.88	.53
B	83	-.67	.03	.95	1.01	.41
B	84	.38	.03	.83	.73	.59
ER	85	2.98	.02	1.02	1.01	.58
ER	86	3.56	.01	1.11	1.09	.54
ER	87	2.60	.02	1.08	1.07	.58

Note: ER = Extended-response item.

Table 3.4
Results of the 2001 Equating Process–Reading Grade 8

Form	Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
AB	1	-1.17	.03	.91	.76	.38
AB	2	1.08	.02	1.14	1.20	.32
AB	3	-.23	.02	1.03	1.02	.36
AB	4	-.49	.02	.94	.85	.41
AB	5	-.42	.02	1.00	.98	.36
AB	6	-.64	.02	.91	.76	.43
AB	7	.61	.02	.95	.91	.48
AB	8	.26	.02	1.02	1.05	.39
AB	9	-.48	.02	.94	1.05	.39
AB	10	.42	.02	1.08	1.11	.35
AB	11	.76	.02	1.02	1.03	.42
AB	12	.50	.02	1.04	1.08	.39
AB	13	.60	.02	1.12	1.21	.32
AB	14	1.42	.02	1.12	1.15	.35
AB	15	-.51	.02	.92	.80	.43
AB	16	.14	.02	.93	.89	.46
AB	17	.57	.02	1.05	1.11	.38
AB	18	.64	.02	1.03	1.05	.40
AB	19	.39	.02	1.00	.97	.43
AB	20	-.26	.02	.93	.94	.43
AB	21	1.45	.02	1.20	1.30	.27
AB	22	-.13	.02	1.09	1.15	.31
AB	23	1.30	.02	1.17	1.25	.30
AB	24	1.02	.02	.98	.98	.46
AB	25	.00	.02	1.03	1.01	.38
AB	26	-.94	.03	.88	.68	.43
AB	27	.84	.02	.95	.93	.48
AB	28	-.69	.02	.97	.92	.37
AB	29	.78	.02	1.20	1.28	.27
AB	30	.04	.02	.93	.84	.46
AB	31	1.19	.02	1.07	1.10	.39
AB	32	-.09	.02	1.01	1.10	.37
AB	33	.15	.02	.92	.86	.47
AB	34	.82	.02	1.06	1.07	.39
AB	35	.24	.02	1.06	1.15	.35
AB	36	2.18	.02	1.09	1.28	.33
AB	37	.51	.02	.93	.88	.48
AB	38	.27	.02	1.00	.99	.41
AB	39	.91	.02	.85	.80	.56
AB	40	1.17	.02	.87	.84	.55
AB	41	-.12	.02	.91	.77	.47
AB	42	1.02	.02	.94	.91	.49
AB	43	.48	.02	1.09	1.13	.35
AB	44	1.32	.02	1.00	1.01	.45
AB	45	-.44	.02	.86	.68	.49
AB	46	-.41	.02	.86	.68	.49
A	47	.98	.03	1.14	1.20	.33
A	48	.17	.03	.99	.91	.41
A	49	1.00	.03	.98	.96	.45

Table 3.4 (continued)

Form	Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
A	50	-.07	.03	1.01	1.01	.41
A	51	1.16	.03	.89	.85	.53
A	52	.50	.03	.99	.96	.44
A	53	.08	.03	.82	.71	.51
A	54	1.47	.03	1.17	1.24	.31
A	55	.71	.03	.95	.92	.46
A	56	.00	.03	1.01	.94	.39
A	57	.79	.03	1.00	1.00	.42
A	58	.80	.03	.98	.95	.42
A	59	.44	.03	.95	.88	.46
A	60	-.32	.03	.93	.81	.48
A	61	1.46	.03	1.05	1.08	.41
A	62	.83	.03	1.01	1.01	.40
A	63	-.23	.03	.88	.79	.48
A	64	-.22	.03	.79	.62	.52
A	65	.41	.03	.96	.90	.44
B	66	.01	.03	1.14	1.34	.26
B	67	.44	.03	1.07	1.11	.36
B	68	.25	.03	1.03	1.02	.39
B	69	.63	.03	.85	.79	.55
B	70	.20	.03	.97	.95	.43
B	71	.25	.03	.95	.94	.45
B	72	1.16	.03	1.05	1.07	.40
B	73	1.21	.03	1.09	1.12	.37
B	74	.99	.03	1.03	1.03	.41
B	75	.62	.03	.94	.88	.49
B	76	.63	.03	.99	.97	.44
B	77	1.10	.03	1.01	1.00	.44
B	78	-.09	.03	.93	.87	.45
B	79	.04	.03	.98	1.12	.40
B	80	.81	.03	.87	.83	.54
B	81	.34	.03	.97	1.03	.43
B	82	1.12	.03	1.02	1.05	.42
B	83	.35	.03	1.04	1.12	.37
B	84	.36	.03	.97	.89	.45
ER	85	2.78	.01	1.02	1.02	.51
ER	86	3.11	.02	1.03	1.03	.51
ER	87	3.13	.02	1.01	1.00	.53

Note: ER = Extended-response item.

Table 3.5
Results of the 2001 Equating Process–Mathematics Grade 3

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
1	-.66	.02	1.02	1.05	.35
2	-.90	.02	.88	.70	.43
3	.04	.02	.92	.88	.48
4	1.22	.02	1.18	1.25	.32
5	.77	.02	.97	.94	.45
6	.90	.02	.96	.93	.49
7	-.28	.02	1.02	1.02	.38
8	-.70	.02	.90	.82	.43
9	-.10	.02	.76	.61	.59
10	.76	.02	1.15	1.19	.34
11	-.36	.02	.87	.84	.52
12	1.17	.02	.97	.97	.48
13	.81	.02	1.27	1.41	.23
14	-.10	.02	.89	.84	.49
15	.72	.02	1.07	1.05	.40
16	-.31	.02	.99	.98	.40
17	.77	.02	1.01	1.03	.44
18	.79	.02	1.04	1.04	.42
19	.58	.02	.97	.95	.49
20	.63	.02	1.06	1.12	.40
21	1.25	.02	1.02	1.04	.44
22	.57	.02	1.08	1.12	.38
23	1.56	.02	.99	.99	.47
24	.62	.02	.96	.95	.48
25	.33	.02	.93	.86	.49
26	.40	.02	.89	.82	.53
27	-.59	.02	1.11	1.24	.33
28	1.33	.02	.98	1.00	.47
29	.29	.02	1.00	.98	.43
30	.14	.02	.84	.76	.50
31	1.40	.02	.86	.87	.56
32	-1.07	.02	.95	.92	.36
33	.33	.02	1.18	1.30	.42
34	.93	.02	1.06	1.07	.41
35	.97	.02	.94	.92	.50
36	-1.50	.03	.93	.78	.35
37	-.69	.02	.82	.67	.44
38	.51	.02	.85	.80	.54
39	.19	.02	.98	.94	.45
40	.58	.02	.95	.95	.47
41	1.83	.02	.95	1.04	.54
42	-.49	.02	.97	.95	.40
43	.74	.02	1.02	1.02	.44
44	1.30	.02	.96	.95	.49
45	.39	.02	.82	.72	.55
46	-.71	.02	.99	.94	.42
47	.87	.02	1.26	1.37	.24
48	.79	.02	.91	.89	.52
49	.31	.02	.85	.78	.54

Table 3.5 (continued)

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
50	1.29	.02	1.01	1.03	.45
51	1.18	.02	1.10	1.14	.38
52	.35	.02	1.16	1.31	.38
53	.25	.02	.87	.82	.54
54	-.24	.02	1.05	1.03	.45
55	1.08	.02	1.26	1.35	.25
56	.75	.02	.91	.87	.51
57	.95	.02	.92	.89	.52
58	.38	.02	.97	.96	.46
59	.42	.02	.98	.93	.44
60	-.30	.02	1.27	1.44	.34
61	.35	.02	.89	.81	.53
62	.12	.02	.84	.81	.53
63	-.84	.02	.74	.59	.38
64	.05	.02	1.03	1.04	.39
65	-.97	.02	1.04	1.31	.28
66	.13	.02	1.01	.96	.43
67	1.07	.02	.97	.95	.49
68	.31	.02	1.19	1.35	.28
69	.22	.02	.92	.87	.49
70	.02	.02	.92	.86	.48
ER1K	.89	.01	1.22	1.24	.64
ER1S	.05	.01	1.18	1.21	.63
ER1E	1.28	.01	1.20	1.24	.59
ER2K	.45	.01	1.71	2.31	.59
ER2S	.35	.01	1.63	1.91	.58
ER2E	1.42	.01	1.27	1.29	.61

Note: ER1, ER2 = Extended-response item; K = Knowledge score; S = Strategy score; E = Explanation score.

Table 3.6
Results of the 2001 Equating Process–Mathematics Grade 5

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
1	-1.92	.03	1.04	.94	.33
2	-.92	.02	.93	.95	.35
3	-.37	.02	1.01	1.02	.37
4	1.05	.02	1.12	1.18	.28
5	-.07	.02	.94	.91	.46
6	.89	.02	.95	.94	.47
7	-1.10	.02	1.05	1.03	.39
8	-.01	.02	1.12	1.18	.29
9	-.59	.02	.96	.93	.40
10	.04	.02	1.01	1.01	.40
11	.51	.02	.95	.94	.47
12	.68	.02	.91	.88	.51
13	-.91	.02	1.13	1.30	.26
14	.10	.02	.97	.95	.43
15	-.03	.02	1.13	1.20	.28
16	.08	.02	.94	.92	.44
17	-.14	.02	.93	.89	.46
18	.34	.02	1.00	.99	.42
19	.92	.02	.95	.96	.48
20	-.13	.02	.81	.74	.59
21	-.08	.02	.92	.88	.46
22	.27	.02	.86	.82	.55
23	.95	.02	1.04	1.06	.39
24	1.32	.02	1.15	1.24	.35
25	.82	.02	1.00	.99	.43
26	.73	.02	1.04	1.06	.39
27	.36	.02	.93	.91	.48
28	.43	.02	.93	.91	.48
29	1.27	.02	.97	.99	.48
30	.36	.02	.88	.84	.54
31	.23	.02	.94	.92	.47
32	.75	.02	.98	.99	.44
33	-.05	.02	.98	.97	.42
34	-.40	.02	.97	.96	.41
35	-.14	.02	.97	.97	.42
36	-1.50	.02	.95	.85	.35
37	-1.72	.03	1.03	1.12	.28
38	.56	.02	.96	.96	.46
39	.27	.02	1.00	.98	.41
40	.92	.02	.92	.91	.50
41	-.50	.02	.86	.74	.51
42	.13	.02	.98	.97	.44
43	.63	.02	1.03	1.06	.39
44	.21	.02	1.06	1.08	.33
45	-.13	.02	1.00	.99	.39
46	1.22	.02	.85	.82	.56
47	.33	.02	1.01	1.01	.42
48	-.72	.02	.94	.86	.52
49	.83	.02	1.09	1.14	.34

Table 3.6 (continued)

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
50	-1.01	.02	.94	.80	.49
51	.20	.02	1.02	1.02	.40
52	-1.48	.02	1.17	1.14	.36
53	.60	.02	1.04	1.06	.39
54	-.70	.02	.96	.92	.40
55	-.71	.02	.79	.66	.55
56	-1.54	.02	.90	.73	.40
57	-1.05	.02	.90	.79	.43
58	.85	.02	.95	.95	.47
59	-.54	.02	.85	.77	.51
60	.38	.02	1.13	1.20	.30
61	1.06	.02	.92	.93	.49
62	.27	.02	.95	.93	.46
63	-.08	.02	.83	.76	.55
64	1.29	.02	1.04	1.09	.38
65	.18	.02	.98	.98	.43
66	.87	.02	.88	.86	.53
67	.63	.02	.83	.80	.57
68	.79	.02	.97	.97	.45
69	.82	.02	.94	.93	.48
70	.78	.02	1.01	1.01	.42
ER1K	-.40	.01	2.06	3.28	.40
ER1S	-.33	.01	2.07	3.40	.40
ER1E	.71	.01	1.61	1.64	.36
ER2K	-.12	.01	1.81	2.08	.41
ER2S	-.17	.01	2.16	3.58	.39
ER2E	.71	.01	1.61	1.64	.36

Note: ER1, ER2 = Extended-response item; K = Knowledge score; S = Strategy score; E = Explanation score.

Table 3.7
Results of the 2001 Equating Process–Mathematics Grade 8

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
1	-.50	.02	1.07	1.31	.31
2	-.63	.02	1.15	1.17	.39
3	.06	.02	.89	.82	.51
4	.47	.02	.94	.90	.50
5	.50	.02	1.00	.94	.49
6	-.43	.02	.86	.73	.50
7	1.14	.02	.93	.91	.53
8	.05	.02	.88	.79	.53
9	-.13	.02	.87	.76	.51
10	.28	.02	1.20	1.33	.30
11	1.34	.02	1.04	1.04	.45
12	.45	.02	.96	.95	.48
13	.98	.02	.99	.97	.49
14	1.93	.02	1.09	1.19	.41
15	1.71	.02	1.01	1.06	.47
16	.28	.02	.87	.81	.54
17	.60	.02	1.02	1.02	.45
18	.81	.02	.89	.85	.55
19	.73	.02	1.04	1.06	.44
20	1.57	.02	1.14	1.23	.38
21	2.05	.02	1.10	1.31	.39
22	2.08	.02	.92	.90	.57
23	1.64	.02	1.20	1.33	.33
24	1.37	.02	1.01	1.03	.47
25	1.07	.02	.85	.80	.59
26	.32	.02	.84	.80	.51
27	.97	.02	.95	.92	.50
28	.08	.02	.93	.83	.49
29	.53	.02	.89	.80	.55
30	-.48	.02	.91	.84	.45
31	.51	.02	.83	.76	.58
32	.15	.02	.90	.82	.51
33	.31	.02	.97	.96	.47
34	.76	.02	.98	.98	.48
35	.80	.02	.94	.91	.52
36	-.95	.02	.92	.75	.41
37	-.60	.02	.96	.83	.42
38	.04	.02	.89	.90	.50
39	.68	.02	1.07	1.07	.42
40	.57	.02	1.08	1.06	.41
41	1.23	.02	1.14	1.17	.38
42	.72	.02	.88	.82	.56
43	.38	.02	1.06	1.07	.45
44	.36	.02	1.04	1.06	.42
45	-.24	.02	.91	.84	.48
46	.53	.02	.99	.98	.47
47	.80	.02	1.14	1.21	.37
48	1.11	.02	.99	.97	.49
49	-.30	.02	.98	.93	.42

Table 3.7 (continued)

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
50	.69	.02	1.13	1.15	.38
51	1.16	.02	1.04	1.04	.46
52	1.19	.02	.86	.82	.58
53	-1.53	.03	1.14	1.16	.36
54	-1.10	.02	1.00	1.03	.37
55	1.14	.02	1.01	1.00	.48
56	.19	.02	.84	.73	.56
57	1.52	.02	.90	.90	.55
58	1.18	.02	1.03	1.03	.46
59	.31	.02	.89	.86	.52
60	1.45	.02	1.04	1.05	.46
61	-.02	.02	.92	.84	.49
62	1.07	.02	.91	.88	.54
63	.30	.02	.91	.87	.51
64	2.00	.02	1.07	1.20	.42
65	-.36	.02	.95	.92	.43
66	.25	.02	.94	.89	.49
67	.45	.02	.98	.99	.47
68	.41	.02	1.00	1.03	.45
69	1.41	.02	1.15	1.27	.37
70	.73	.02	.89	.85	.55
ER1K	.37	.01	1.63	1.88	.54
ER1S	.39	.01	1.86	2.21	.52
ER1E	.90	.01	1.61	1.69	.44
ER2K	.41	.01	1.20	1.28	.62
ER2S	.21	.01	1.39	1.70	.60
ER2E	.90	.01	1.38	1.47	.54

Note: ER1, ER2 = Extended-response item; K = Knowledge score; S = Strategy score; E = Explanation score.

Table 3.8
Results of the 2001 Scaling Process–Science Grade 4

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
1	-.61	.02	.97	.94	.38
2	-.78	.02	.93	.92	.32
3	-.82	.02	1.01	1.03	.34
4	-.45	.02	1.10	1.16	.27
5	.32	.02	.95	.93	.44
6	1.98	.02	1.15	1.56	.10
7	.73	.02	1.20	1.26	.19
8	.20	.02	1.15	1.19	.24
9	-.30	.02	.91	.89	.42
10	.10	.02	.94	.91	.43
11	-1.02	.02	.97	.96	.32
12	-1.10	.02	1.00	1.03	.32
13	-1.05	.02	.98	1.00	.35
14	-1.04	.02	1.00	.97	.34
15	.25	.02	1.07	1.11	.32
16	.29	.02	1.07	1.07	.31
17	.82	.02	1.16	1.21	.24
18	.58	.02	1.10	1.12	.30
19	-.39	.02	.94	.91	.41
20	.34	.02	.92	.89	.47
21	-.13	.02	.98	.95	.40
22	.15	.02	.89	.85	.50
23	.49	.02	1.15	1.17	.25
24	-.99	.02	.84	.73	.50
25	.39	.02	1.06	1.06	.33
26	.67	.02	1.11	1.15	.27
27	.06	.02	.97	.96	.43
28	-.87	.02	.89	.83	.47
29	.14	.02	.93	.90	.46
30	-.31	.02	.84	.77	.55
31	.88	.02	1.05	1.06	.33
32	.68	.02	1.11	1.14	.30
33	-.11	.02	.95	.92	.49
34	-.10	.02	.99	.98	.40
35	-.26	.02	.94	.90	.44
41	-1.07	.02	.93	.88	.41
42	-.44	.02	1.07	1.11	.33
43	1.56	.02	1.19	1.38	.15
44	.03	.02	1.08	1.11	.32
45	-1.25	.02	1.07	1.00	.44
46	.18	.02	1.10	1.14	.30
47	.26	.02	.95	.93	.44
48	.09	.02	.88	.84	.51
49	-.97	.02	.92	.82	.43
50	.45	.02	1.13	1.17	.27
51	-.85	.02	.88	.79	.51
52	-.03	.02	.90	.85	.49
53	-.19	.02	1.06	1.12	.32
54	-.37	.02	1.02	1.03	.35

Table 3.8 (continued)

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
55	.14	.02	.98	.96	.41
56	-.30	.02	1.12	1.22	.23
57	.11	.02	.93	.93	.46
58	-.54	.02	.88	.81	.50
59	.72	.02	1.05	1.08	.33
60	.02	.02	.95	.94	.45
61	.25	.02	.97	.96	.41
62	-.36	.02	.93	.87	.45
63	-.53	.02	.83	.73	.55
64	-.10	.02	.83	.77	.52
65	-.43	.02	.95	.92	.42
66	.04	.02	.92	.89	.48
67	.45	.02	1.10	1.11	.29
68	.65	.02	1.26	1.33	.15
69	.73	.02	1.08	1.10	.30
70	-.10	.02	.86	.81	.55
71	-.78	.02	.91	.85	.43
72	.63	.02	1.00	1.01	.39
73	.37	.02	.90	.88	.49
74	-.01	.02	1.04	1.05	.37
75	.31	.02	1.08	1.09	.32

Table 3.9
Results of the 2001 Scaling Process–Science Grade 7

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
1	.73	.02	1.13	1.18	.24
2	-1.51	.02	.91	.78	.39
3	-.43	.02	.92	.83	.45
4	-.67	.02	.94	.88	.40
5	-1.07	.02	.93	.79	.46
6	-.73	.02	.92	.84	.41
7	.38	.02	.93	.90	.44
8	-.84	.02	1.00	1.00	.32
9	-.69	.02	.84	.74	.45
10	-.86	.02	1.03	1.10	.28
11	.83	.02	1.06	1.10	.31
12	-.39	.02	.95	.94	.40
13	-1.60	.02	1.04	1.39	.18
14	-.17	.02	.89	.83	.49
15	.06	.02	1.00	.99	.37
16	-.31	.02	.91	.87	.39
17	-.71	.02	1.04	1.06	.29
18	-.09	.02	1.02	1.02	.34
19	-.14	.02	.95	.96	.41
20	-.36	.02	.92	.88	.44
21	-.87	.02	.91	.84	.43
22	-.03	.02	1.06	1.10	.31
23	1.23	.02	.99	1.01	.37
24	-.03	.02	.95	.91	.43
25	-.95	.02	1.03	1.04	.32
26	.67	.02	1.04	1.05	.33
27	-.18	.02	.85	.79	.53
28	.86	.02	1.03	1.06	.33
29	.48	.02	1.05	1.09	.33
30	.21	.02	.98	.97	.41
31	-.33	.02	.86	.78	.51
32	.74	.02	.96	.97	.40
33	-.48	.02	1.04	1.12	.30
34	.21	.02	1.08	1.10	.29
35	.20	.02	1.02	1.03	.36
41	-.07	.02	1.09	1.11	.27
42	-.14	.02	.89	.84	.48
43	.32	.02	1.08	1.10	.30
44	-.25	.02	.97	.92	.46
45	-.87	.02	.94	.90	.39
46	-1.46	.02	.97	1.07	.29
47	.44	.02	.96	.94	.41
48	.40	.02	1.17	1.21	.20
49	-.95	.02	.99	1.05	.32
50	-.61	.02	.88	.79	.47
51	.09	.02	.98	.96	.37
52	.66	.02	1.13	1.17	.25
53	1.21	.02	1.25	1.37	.11
54	.81	.02	1.05	1.08	.32

Table 3.9 (continued)

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
55	-.57	.02	.93	.92	.42
56	-.39	.02	1.16	1.27	.26
57	-.16	.02	1.07	1.07	.35
58	.04	.02	1.00	.99	.37
59	.67	.02	.98	.98	.39
60	.35	.02	1.15	1.21	.23
61	-.29	.02	.99	.99	.31
62	-.27	.02	.96	.98	.42
63	.23	.02	.97	.94	.41
64	.69	.02	1.19	1.23	.19
65	.50	.02	1.08	1.09	.30
66	.32	.02	.95	.93	.42
67	-.87	.02	.93	.87	.41
68	.90	.02	1.10	1.14	.27
69	-.42	.02	.98	.95	.38
70	-.40	.02	.88	.80	.48
71	-.46	.02	.89	.84	.47
72	1.10	.02	.96	.98	.39
73	.57	.02	1.17	1.22	.20
74	.46	.02	1.07	1.08	.30
75	.13	.02	.95	.93	.42

Table 3.10
Results of the 2001 Scaling Process–Social Science Grade 4

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
1	-1.08	.02	.91	.81	.42
2	-.78	.02	.86	.73	.50
3	.47	.02	.96	.94	.44
4	-.69	.02	1.11	1.22	.34
5	-.51	.02	.94	.87	.44
6	.07	.02	1.15	1.21	.25
7	-.63	.02	1.18	1.29	.31
8	.07	.02	1.13	1.22	.26
9	-.02	.02	.91	.86	.45
10	-.72	.02	1.02	1.11	.32
11	-.87	.02	1.06	1.21	.26
12	2.04	.02	1.14	1.50	.16
13	-.02	.02	.88	.80	.49
14	.62	.02	1.04	1.09	.36
15	-1.10	.02	.85	.69	.47
16	-.38	.02	.94	.90	.45
17	1.30	.02	1.07	1.15	.30
18	-.86	.02	.93	.97	.40
19	.18	.02	.95	.94	.44
20	-.86	.02	.84	.71	.49
21	-.50	.02	.86	.79	.51
22	.96	.02	1.14	1.20	.27
23	-.12	.02	.92	.88	.43
24	.28	.02	1.04	1.10	.36
25	1.51	.02	1.05	1.14	.31
26	-.16	.02	.95	.92	.48
27	-.02	.02	1.12	1.13	.27
28	.07	.02	1.10	1.14	.29
29	.61	.02	1.08	1.11	.32
30	.46	.02	1.08	1.07	.32
31	1.39	.02	1.14	1.28	.25
32	1.03	.02	.98	1.00	.42
33	-1.14	.02	.90	.92	.41
34	.26	.02	.96	.95	.43
35	.54	.02	1.26	1.34	.15
36	-.02	.02	.84	.77	.55
37	.82	.02	1.05	1.08	.34
38	-.27	.02	.76	.66	.59
39	.95	.02	1.01	1.02	.41
40	.40	.02	.94	.91	.45
41	-.21	.02	.93	.89	.45
42	-.77	.02	.87	.77	.48
43	.42	.02	1.05	1.07	.35
44	1.03	.02	1.12	1.19	.27
45	-.29	.02	1.10	1.15	.28
46	-.33	.02	1.12	1.14	.31
47	.17	.02	1.14	1.16	.27
48	.32	.02	1.06	1.06	.36
49	-1.65	.03	.90	.78	.38

Table 3.10 (continued)

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
50	-.19	.02	.99	1.00	.39
51	-1.34	.02	.91	.80	.40
52	-.28	.02	1.08	1.16	.29
53	1.53	.02	1.17	1.33	.20
54	-.78	.02	.88	.85	.47
55	.00	.02	.98	.96	.41
56	.09	.02	1.00	.97	.39
57	.68	.02	.96	.96	.43
58	.40	.02	.99	.98	.40
59	-.61	.02	.94	.86	.42
60	.70	.02	.98	.97	.41
61	-.17	.02	.94	.93	.41
62	-1.63	.03	.92	.85	.36
63	.07	.02	.93	.92	.48
64	.38	.02	.95	.92	.43
65	.52	.02	1.12	1.16	.28
66	-.29	.02	1.00	.96	.42
67	.11	.02	.91	.86	.48
68	.62	.02	.99	.99	.41
69	.57	.02	.97	.97	.42
70	1.40	.02	1.11	1.27	.24

Table 3.11
Results of the 2001 Scaling Process–Social Science Grade 7

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
1	-1.53	.03	.99	1.13	.23
2	-.11	.02	1.14	1.20	.17
3	-.43	.02	.90	.82	.44
4	.09	.02	1.03	1.01	.35
5	.55	.02	1.04	1.04	.33
6	-1.26	.02	1.06	1.05	.29
7	1.01	.02	1.03	1.05	.33
8	.25	.02	1.07	1.08	.27
9	.78	.02	1.11	1.12	.24
10	-.14	.02	1.01	1.05	.32
11	-.49	.02	.93	.86	.40
12	.50	.02	.95	.93	.41
13	1.41	.02	1.08	1.14	.26
14	.14	.02	1.06	1.08	.27
15	.78	.02	1.11	1.14	.23
16	.39	.02	1.10	1.13	.24
17	-.39	.02	.89	.82	.42
18	-.69	.02	.96	.94	.34
19	.28	.02	1.07	1.05	.28
20	.36	.02	.94	.96	.42
21	1.16	.02	1.10	1.16	.23
22	.56	.02	1.03	1.04	.33
23	.76	.02	1.05	1.06	.30
24	-1.08	.02	.88	.72	.43
25	-.07	.02	.94	.93	.35
26	.40	.02	1.00	1.02	.35
27	1.27	.02	1.05	1.09	.28
28	.54	.02	.96	.94	.40
29	1.32	.02	1.13	1.17	.25
30	.58	.02	.98	.99	.37
31	.88	.02	1.18	1.23	.16
32	-.51	.02	1.04	1.09	.26
33	-.77	.02	.94	.89	.37
34	1.26	.02	1.19	1.25	.15
35	.51	.02	1.01	1.00	.34
36	-.03	.02	.94	.89	.42
37	.45	.02	.89	.85	.47
38	1.01	.02	1.07	1.10	.28
39	.69	.02	.89	.87	.48
40	.50	.02	.94	.92	.42
41	.48	.02	.94	.92	.42
42	.04	.02	1.00	.98	.35
43	.96	.02	1.01	1.02	.34
44	-.04	.02	.96	.92	.34
45	-.39	.02	.90	.84	.44
46	.69	.02	.98	.99	.38
47	-1.25	.02	1.05	.98	.34
48	-.46	.02	.90	.83	.43
49	-.41	.02	.90	.85	.44

Table 3.11 (continued)

Item	Difficulty	S_{ed}	Infit	Outfit	r_{pb}
50	.50	.02	.99	.97	.37
51	.68	.02	1.11	1.13	.24
52	1.22	.02	1.20	1.27	.18
53	-.09	.02	.99	.98	.37
54	.98	.02	1.02	1.02	.33
55	.89	.02	1.06	1.08	.29
56	-.18	.02	.96	.92	.42
57	.29	.02	.96	.93	.41
58	-.57	.02	.99	.91	.47
59	-.45	.02	.97	.97	.36
60	-1.40	.02	.94	.90	.32
61	.55	.02	1.04	1.06	.31
62	.81	.02	.97	.97	.38
63	-.05	.02	.89	.83	.47
64	1.14	.02	1.07	1.11	.27
65	.34	.02	.90	.86	.46
66	.02	.02	.91	.86	.45
67	.59	.02	.94	.92	.42
68	.53	.02	.94	.93	.42
69	1.12	.02	1.08	1.12	.26
70	1.27	.02	1.08	1.13	.25

The raw score that is initially derived from multiple-choice items 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. RESULTS

Performance Relative to the Illinois Learning Standards

Table 4.1 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. 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 4.1
Percentages of Students by Grade Falling into Each Performance Level for ISAT Reading: 1999-2001

Grade/ Year	Academic Warning	Below Standards	Meets Standards	Exceeds Standards
3				
1999	8	31	44	17
2000	6	32	41	21
2001	7	31	43	19
5				
1999	1	38	37	24
2000	0	41	39	20
2001	1	40	34	25
8				
1999	1	27	54	18
2000	0	28	56	16
2001	1	34	56	10

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

Table 4.2 provides additional information with respect to the reading test. It presents the average percent of items students answered correctly with respect to the standard sets that were previously described.

Table 4.2
Reading Average Percent Correct by Standard Sets: 2001

	Set					
Grade	1	2	3	4	5	6
03	68	69	66	66	74	71
05	64	66	66	67	63	–
08	71	66	70	66	69	–

Table 4.3 shows the percentages of students by performance level and by grade for mathematics. The percentage of students meeting state standards is highest for grade 3

students and lowest for grade 8 students. The percentage of students falling into the Exceeds category is much higher at grade 3 than at the other two grades.

Table 4.3
Percentages of Students by Grade Falling into Each Performance Level for ISAT Mathematics: 1999-2001

Grade/ Year	Academic Warning	Below Standards	Meets Standards	Exceeds Standards
3				
1999	12	20	47	21
2000	10	21	46	23
2001	8	18	46	28
5				
1999	6	39	53	3
2000	6	37	52	5
2001	4	34	55	6
8				
1999	5	52	36	7
2000	8	46	35	12
2001	7	42	37	13

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

Table 4.4 presents the average percent of items students answered correctly with respect to the mathematics standard sets that were previously described.

Table 4.4
Mathematics Average Percent Correct by Standard Sets: 2001

	Set							
Grade	1	2	3	4	5	6	7	8
03	63	64	62	63	74	55	71	63
05	57	59	56	61	63	56	57	64
08	58	55	56	56	54	56	62	59

Table 4.5 shows results for writing. A greater percentage of 5th- and 8th-grade students meet standards with respect to writing as compared to 3rd-graders.

Table 4.6 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 4.5
Percentages of Students by Grade Falling into Each Performance Level for ISAT Writing: 1999-2001

Grade/ Year	Academic Warning	Below Standards	Meets Standards	Exceeds Standards
3				
1999	9	35	50	6
2000	6	38	53	2
2001	9	33	55	3
5				
1999	2	23	52	23
2000	3	26	57	14
2001	4	27	58	12
8				
1999	5	36	56	3
2000	3	27	59	11
2001	6	32	55	7

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

Table 4.6
Mean Writing Feature Scores of Students by Prompt: 2001

Grade	Type	F	S	O	C	I
03	P	4.5	3.7	3.5	1.9	3.7
03	E	4.5	3.8	3.6	1.9	3.8
03	N	3.7	4.0	3.7	1.9	3.7
05	P	4.8	3.9	3.8	2.0	3.9
05	E	4.6	3.8	3.8	2.0	3.8
05	N	4.0	4.1	3.9	2.0	4.0
08	P	3.9	3.6	3.7	1.9	3.7
08	E	3.6	3.4	3.5	1.9	3.5
08	N	4.1	4.1	4.0	1.9	4.1

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

Table 4.7 shows the percentages of students by performance level and by grade for science.

Table 4.7
Percentages of Students by Grade Falling into Each Performance Level for ISAT Science: 2000-2001

Grade/ Year	Academic Warning	Below Standards	Meets Standards	Exceeds Standards
4				
2000	1	35	51	13
2001	8	26	54	11
7				
2000	12	16	54	18
2001	11	17	52	20

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

Table 4.8 presents the average percent of items students answered correctly with respect to the science standards sets that were previously described.

Table 4.8
Science Average Percent Correct by Standard Sets: 2001

	Set				
Grade	1	2	3	4	5
04	58	60	61	62	60
07	64	65	65	61	62

Table 4.9 shows the percentages of students by performance level and by grade for social science.

Table 4.9
Percentages of Students by Grade Falling into Each Performance Level for ISAT Social Science: 2000-2001

Grade/ Year	Academic Warning	Below Standards	Meets Standards	Exceeds Standards
04				
2000	11	30	53	6
2001	11	28	55	6
07				
2000	3	39	46	12
2001	2	38	47	13

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

Table 4.10 presents the average percent of items students answered correctly with respect to the social science standard sets that were previously described.

Table 4.10
Social Science Average Percent Correct by Standard Sets: 2001

	Set				
Grade	1	2	3	4	5
04	61	65	60	59	61
07	60	59	57	58	59

Performance Relative to National Quarters

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. National norms for writing are not provided because no nationally standardized writing test has a sufficiently satisfactory match to the Illinois content specifications.

ISAT uses the Ninth Edition of the Stanford Achievement Tests (SAT9) for purposes of determining Illinois students' relative standing within the national population. 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.

Table 4.11 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 145 on the 4th-grade science 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.

The results of applying these cutoffs to the 2001 assessment data are shown in Table 4.12. As noted earlier, results in writing are not reported relative to national quarters.

Table 4.11
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
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
SCIENCE	**Q1**	**Q2**	**Q3**	**Q4**
04	120-145	146-157	158-168	169-200
07	120-142	143-154	155-163	164-200
SOCIAL SCIENCE	**Q1**	**Q2**	**Q3**	**Q4**
04	120-144	145-155	156-166	167-200
07	120-145	146-154	155-165	166-200

Table 4.12
Percentages of Students by Grade and Learning Area Falling into Each National Quartile: 1999-2001

READING	Q1	Q2	Q3	Q4
Grade/Year				
3				
1999	22	22	25	32
2000	21	21	25	33
2001	21	22	25	32
5				
1999	21	23	27	28
2000	21	26	28	25
2001	25	21	24	30
8				
1999	15	22	30	33
2000	13	24	33	30
2001	17	26	33	24

MATHEMATICS	Q1	Q2	Q3	Q4
Grade/Year				
3				
1999	19	21	28	32
2000	18	21	26	36
2001	14	19	25	42
5				
1999	20	22	24	33
2000	19	22	21	38
2001	17	19	21	42
8				
1999	15	25	25	35
2000	18	20	21	41
2001	17	19	18	45

SCIENCE	Q1	Q2	Q3	Q4
Grade/Year				
4				
2000	18	26	25	31
2001	19	23	27	30
7				
2000	14	24	22	41
2001	12	25	20	43

SOCIAL SCIENCE	Q1	Q2	Q3	Q4
Grade/Year				
4				
2000	17	21	29	33
2001	16	21	28	35
7				
2000	17	19	29	35
2001	16	18	27	38

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

Correlations Among Scale Scores

Correlations among the scale scores at each grade tested are presented in Table 4.13. Appendix A provides correlations among the standard sets as well as breakdowns by writing genre. The sample sizes on which the correlations in Table 4.13 are based are also shown in Appendix A.

Table 4.13
Correlations Among ISAT Scale Scores

Grade 3	Reading	Mathematics	Writing
Reading	1.000	.806	.437
Mathematics	.806	1.000	.419
Writing	.437	.419	1.000

Grade 5	Reading	Mathematics	Writing
Reading	1.000	.794	.575
Mathematics	.794	1.000	.575
Writing	.575	.575	1.000

Grade 8	Reading	Mathematics	Writing
Reading	1.000	.788	.615
Mathematics	.788	1.000	.581
Writing	.615	.581	1.000

Grade 4	Science	Social Science
Science	1.000	.867
Social Science	.867	1.000

Grade 7	Science	Social Science
Science	1.000	.858
Social Science	.858	1.000

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 standard 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 and 8, the sample size refers to the number of papers, not the number of students.

Reading: Grade 3	142,996
Reading: Grade 5	153,230
Reading: Grade 8	139,551
Mathematics: Grade 3	143,716
Mathematics: Grade 5	153,725
Mathematics: Grade 8	140,166
Writing: Persuasive Prompt: Grade 3	47,134
Writing: Expository Prompt: Grade 3	47,301
Writing: Narrative Prompt: Grade 3	46,137
Writing: Persuasive Prompt: Grade 5	116,012
Writing: Expository Prompt: Grade 5	36,355
Writing: Narrative Prompt: Grade 5	152,354
Writing: Persuasive Prompt: Grade 8	119,370
Writing: Expository Prompt: Grade 8	19,506
Writing: Narrative Prompt: Grade 8	139,277
Science: Grade 4	153,028
Science: Grade 7	143,631
Social Science: Grade 4	153,279
Social Science: Grade 7	143,741

Table A.1
Correlations Among Reading Standard Sets

Grade 3	S1	S2	S3	S4	S5	S6
S1	1.000	.819	.832	.923	.851	.601
S2	.825	1.000	.942	.860	.900	.653
S3	.838	.939	1.000	.799	.800	.621
S4	.922	.864	.796	1.000	.790	.593
S5	.858	.903	.799	.800	1.000	.632
S6	.598	.654	.614	.598	.631	1.000

Grade 5	S1	S2	S3	S4	S5
S1	1.000	.813	.873	.877	.791
S2	.819	1.000	.905	.952	.892
S3	.886	.921	1.000	.847	.788
S4	.875	.954	.865	1.000	.811
S5	.798	.889	.800	.815	1.000

Grade 8	S1	S2	S3	S4	S5
S1	1.000	.765	.731	.897	.752
S2	.744	1.000	.933	.917	.887
S3	.700	.914	1.000	.809	.775
S4	.880	.934	.806	1.000	.783
S5	.724	.865	.743	.777	1.000

Note: Values for Form A are presented above the principal diagonal, and values for Form B are presented below the principal diagonal.

Table A.2
Correlations Among Mathematics Standard Sets

Grade 3	S1	S2	S3	S4	S5	S6	S7	S8
S1	1.000	.875	.766	.796	.687	.843	.866	.843
S2	.875	1.000	.722	.692	.618	.721	.746	.767
S3	.766	.722	1.000	.653	.585	.647	.684	.658
S4	.796	.692	.653	1.000	.782	.709	.711	.734
S5	.687	.618	.585	.782	1.000	.596	.647	.619
S6	.843	.721	.647	.709	.596	1.000	.684	.689
S7	.866	.746	.684	.711	.647	.684	1.000	.726
S8	.843	.767	.658	.734	.619	.689	.726	1.000

Grade 5	S1	S2	S3	S4	S5	S6	S7	S8
S1	1.000	.938	.827	.801	.797	.893	.848	.843
S2	.938	1.000	.800	.743	.766	.780	.743	.772
S3	.827	.800	1.000	.668	.657	.720	.667	.686
S4	.801	.743	.668	1.000	.764	.742	.674	.692
S5	.797	.766	.657	.764	1.000	.697	.663	.683
S6	.893	.780	.720	.742	.697	1.000	.700	.712
S7	.848	.743	.667	.674	.663	.700	1.000	.690
S8	.843	.772	.686	.692	.683	.712	.690	1.000

Grade 8	S1	S2	S3	S4	S5	S6	S7	S8
S1	1.000	.944	.837	.850	.837	.879	.864	.876
S2	.944	1.000	.842	.795	.809	.862	.777	.801
S3	.837	.842	1.000	.757	.759	.735	.745	.734
S4	.850	.795	.757	1.000	.825	.727	.721	.731
S5	.837	.809	.759	.825	1.000	.746	.721	.732
S6	.879	.862	.735	.727	.746	1.000	.740	.766
S7	.864	.777	.745	.721	.721	.740	1.000	.783
S8	.876	.801	.734	.731	.732	.766	.783	1.000

Table A.3
Correlations Among Writing Feature Scores

Persuasive Prompt: Grade 3	F	S	O	C	I
F	1.000	.740	.787	.423	.799
S	.740	1.000	.880	.381	.949
O	.787	.880	1.000	.394	.929
C	.423	.381	.394	1.000	.394
I	.799	.949	.929	.394	1.000

Expository Prompt: Grade 3	F	S	O	C	I
F	1.000	.702	.766	.425	.791
S	.702	1.000	.858	.391	.935
O	.766	.858	1.000	.403	.916
C	.425	.391	.403	1.000	.406
I	.791	.935	.916	.406	1.000

Narrative Prompt: Grade 3	F	S	O	C	I
F	1.000	.850	.963	.347	.968
S	.850	1.000	.860	.371	.890
O	.963	.860	1.000	.354	.977
C	.347	.371	.354	1.000	.352
I	.968	.890	.977	.352	1.000

Persuasive Prompt: Grade 5	F	S	O	C	I
F	1.000	.705	.713	.345	.719
S	.705	1.000	.944	.289	.977
O	.713	.944	1.000	.301	.966
C	.345	.289	.301	1.000	.298
I	.719	.977	.966	.298	1.000

Expository Prompt: Grade 5	F	S	O	C	I
F	1.000	.763	.769	.417	.777
S	.763	1.000	.957	.358	.984
O	.769	.957	1.000	.366	.972
C	.417	.358	.366	1.000	.367
I	.777	.984	.972	.367	1.000

Table A.3 (continued)

Narrative Prompt: Grade 5	F	S	O	C	I
F	1.000	.819	.930	.246	.953
S	.819	1.000	.830	.258	.883
O	.930	.830	1.000	.247	.955
C	.246	.258	.247	1.000	.259
I	.953	.883	.955	.259	1.000

Persuasive Prompt: Grade 8	F	S	O	C	I
F	1.000	.776	.791	.403	.818
S	.776	1.000	.935	.401	.954
O	.791	.935	1.000	.414	.973
C	.403	.401	.414	1.000	.418
I	.818	.954	.973	.418	1.000

Expository Prompt: Grade 8	F	S	O	C	I
F	1.000	.854	.871	.518	.886
S	.854	1.000	.955	.499	.966
O	.871	.955	1.000	.513	.986
C	.518	.499	.513	1.000	.516
I	.886	.966	.986	.516	1.000

Narrative Prompt: Grade 8	F	S	O	C	I
F	1.000	.887	.894	.331	.918
S	.887	1.000	.938	.325	.969
O	.894	.938	1.000	.334	.968
C	.331	.325	.334	1.000	.332
I	.918	.969	.968	.332	1.000

Table A.4
Correlations Among Science Standard Sets

Grade 4	S1	S2	S3	S4	S5
S1	1.000	.684	.667	.692	.694
S2	.684	1.000	.685	.709	.708
S3	.667	.685	1.000	.698	.693
S4	.692	.709	.698	1.000	.726
S5	.694	.708	.693	.726	1.000

Grade 7	S1	S2	S3	S4	S5
S1	1.000	.709	.670	.679	.692
S2	.709	1.000	.648	.676	.695
S3	.670	.648	1.000	.645	.628
S4	.679	.676	.645	1.000	.676
S5	.692	.695	.628	.676	1.000

Table A.5
Correlations Among Social Science Standard Sets

Grade 4	S1	S2	S3	S4	S5
S1	1.000	.677	.676	.665	.689
S2	.677	1.000	.693	.653	.676
S3	.676	.693	1.000	.671	.694
S4	.665	.653	.671	1.000	.679
S5	.689	.676	.694	.679	1.000

Grade 7	S1	S2	S3	S4	S5
S1	1.000	.631	.616	.621	.656
S2	.631	1.000	.622	.618	.656
S3	.616	.622	1.000	.621	.683
S4	.621	.618	.621	1.000	.652
S5	.656	.656	.683	.652	1.000

APPENDIX B. PREDICTING 5TH-GRADE ISAT PERFORMANCE IN READING, MATHEMATICS, AND WRITING FROM 3RD-GRADE ISAT SCORES

In 2001, many 5th-grade students who took ISAT had participated in the first administration of ISAT tests of reading, mathematics, and writing in 1999 as third-graders.

Although Illinois does not use a common identification number to uniquely identify students in the state, it is possible to match records across years using local identification codes. A match of the 1999 and 2001 data sets by this criterion identified approximately 72,000 students from 627 districts across the state for whom it was possible to match grade 3 results with grade 5 results.

Table B.1
Correlations Among ISAT Tests

	Reading 3	Mathematics 3	Writing 3	Reading 5	Mathematics 5	Writing 5
Reading 3	1.000					
Mathematics 3	.739	1.000				
Writing 3	.487	.475	1.000			
Reading 5	**.780**	.689	.451	1.000		
Mathematics 5	.710	**.809**	.461	.775	1.000	
Writing 5	.494	.467	**.444**	.545	.547	1.000

Correlations among the tests are shown in Table B.1. The correlation between grade 3 and grade 5 reading scores is .78. Comparable values are .81 for mathematics and .44 for writing. The lower value for writing is to be expected, both because of the performance nature of the ISAT writing test and because 3rd-grade students write only to a single prompt.

It is also possible to examine the relationship between categorical performance on the tests. For example, the data in Table B.2 shows the relationship between students' classifications on the reading tests. Rows of the table represent the 3rd-grade outcome, and columns represent the 5th-grade outcome. The first set of numbers shows the actual count of students. The second set of numbers shows the percentage of the total sample falling into each cell.

For example, 1,420 students (2% of the total data set) were classified as Academic Warning both times. Overall, 63% of students in the study were identically classified in both grades. Approximately 20% of the total sample were classified at a higher level in 5th grade than they had been in 3rd grade, and 18% of the sample were classified at a lower level in 5th grade than they had been earlier. Most of the shifts in categories from one grade to another are adjacent (i.e., Meets to Exceeds). Less than one percent of students shifted more than one category during the two years.

Table B.2
Relationship Between Performance Classifications in Reading

Grade 3 Reading Classification	Grade 5 Reading Classification (N)			
	Warning	Below	Meet	Exceed
Warning	1420	1841	142	2
Below	1836	12859	5632	218
Meet	155	5883	21374	6486
Exceed	8	168	4533	9232

Grade 3 Reading Classification	Grade 5 Reading Classification (%)			
	Warning	Below	Meet	Exceed
Warning	2.0%	2.6%	0.2%	0.0%
Below	2.6%	17.9%	7.8%	0.3%
Meet	0.2%	8.2%	29.8%	9.0%
Exceed	0.0%	0.2%	6.3%	12.9%

Table B.3 presents parallel information on the relationship between mathematics performance classifications. The outcomes are similar. Overall, 62% of students in the study were identically classified in both grades. The shift upward was more dramatic, however, for mathematics than for reading. Approximately 28% of the total sample were classified at a higher level in 5th grade than they had been in 3rd grade. Only 10% of the sample were classified at a lower level in 5th grade than they had been earlier. Most of the shifts in categories from one grade to another are adjacent. But a slightly higher percentage (2%) of students shifted more than one category during the two years.

Table B.3
Relationship Between Performance Classifications in Mathematics

Grade 3 Mathematics Classification	Grade 5 Mathematics Classification (N)			
	Warning	Below	Meet	Exceed
Warning	2192	2563	1057	25
Below	1374	4972	6380	270
Meet	318	2805	22227	10052
Exceed	8	53	2590	14644

Grade 3 Mathematics Classification	Grade 5 Mathematics Classification (%)			
	Warning	Below	Meet	Exceed
Warning	3.1%	3.6%	1.5%	0.0%
Below	1.9%	7.0%	8.9%	0.4%
Meet	0.4%	3.9%	31.1%	14.1%
Exceed	0.0%	0.1%	3.6%	20.5%

Table B.4 presents parallel information on the relationship between writing performance classifications. The outcomes are unlike those in reading and opposite to the findings for

mathematics. Overall, 53% of students in the study were identically classified in both grades. The shift downward was more dramatic, particularly when compared to the results for mathematics. Approximately 29% of the total sample were classified at a lower performance level in 5th grade than they had been in 3rd grade. Only 18% of the sample were classified at a higher level in 5th grade than they had been earlier. As with the other two areas, most of the shifts in categories from one grade to another are adjacent. But an even higher percentage (4%) of students shifted more than one category during the two years.

Table B.4
Relationship Between Performance Classifications in Writing

Grade 3 Writing	Grade 5 Writing Classification (N)			
Classification	Warning	Below	Meet	Exceed
Warning	479	2690	1430	18
Below	530	9809	12530	561
Meet	150	8086	25857	3266
Exceed	6	405	3277	1097
Grade 3 Writing	**Grade 5 Writing Classification (%)**			
Classification	Warning	Below	Meet	Exceed
Warning	0.7%	3.8%	2.0%	0.0%
Below	0.8%	14.0%	17.9%	0.8%
Meet	0.2%	11.5%	36.8%	4.7%
Exceed	0.0%	0.6%	4.7%	1.6%

Figures B.1 through B.3 present a second way of looking at the data of these tables. These figures show the probability of meeting or exceeding the 5th-grade standard associated with each 3rd-grade scale score value. Scores on the 3rd-grade tests are represented by the horizontal axis. The vertical axis shows the probability of meeting or exceeding the 5th-grade standard. With respect to reading, for example, a scale score of 140 is associated with a probability of only about .10 of meeting the 5th-grade standard. At the other end of the scale, students who scored 170 or higher on the 3rd-grade test have a very high probability (.90) of meeting or exceeding the 5th-grade standard. A score of 156, which is the minimum required to be classified as meeting standards on the 3rd-grade test, is associated with a .56 probability of meeting or exceeding the 5th-grade standard (i.e., obtaining a score of 156 or higher on the 5th-grade test).

Figure B.1
Relationship Between ISAT Reading Tests

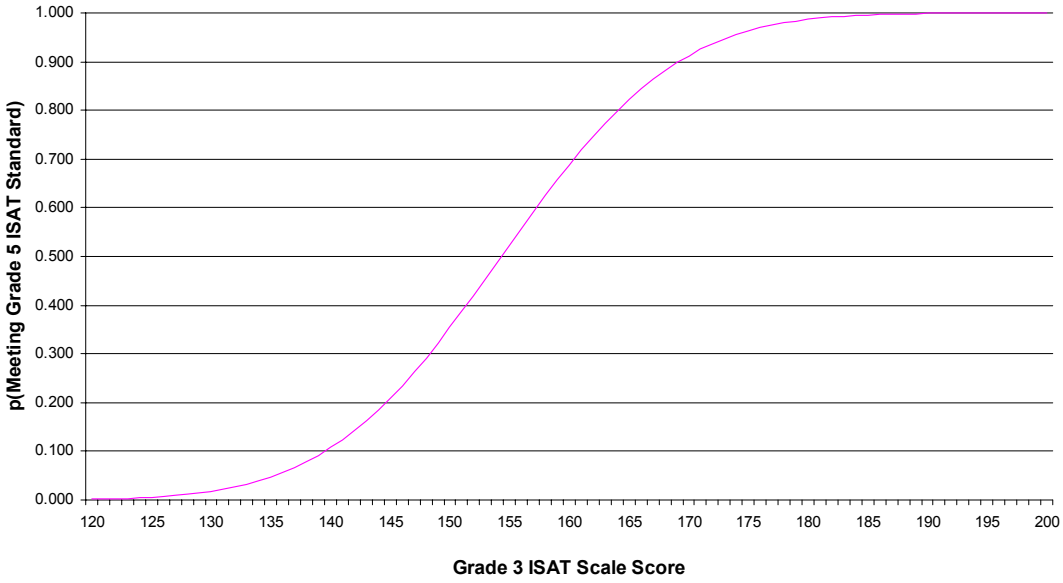


Figure B.2
Relationship Between ISAT Mathematics Tests

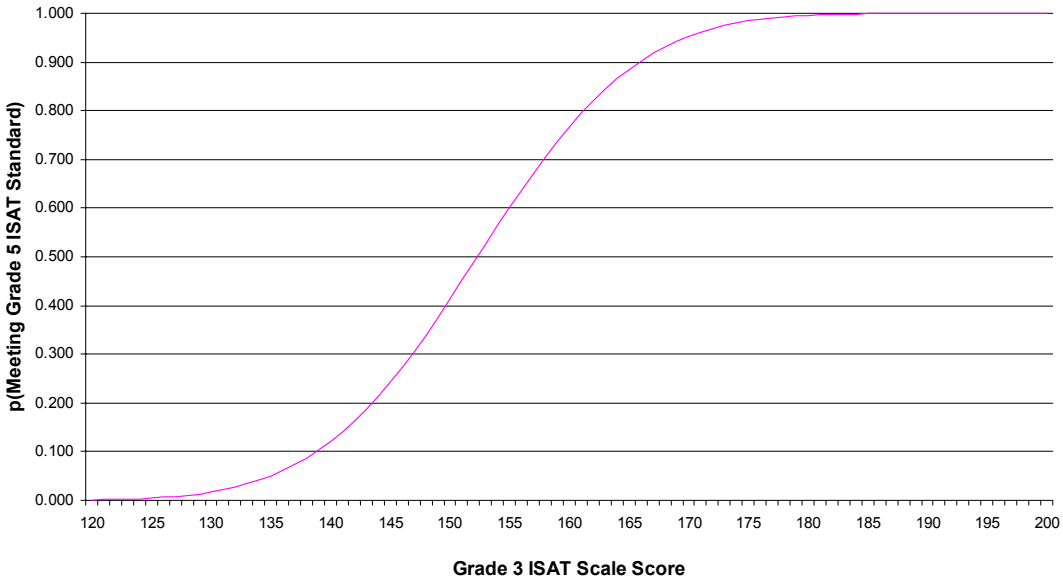
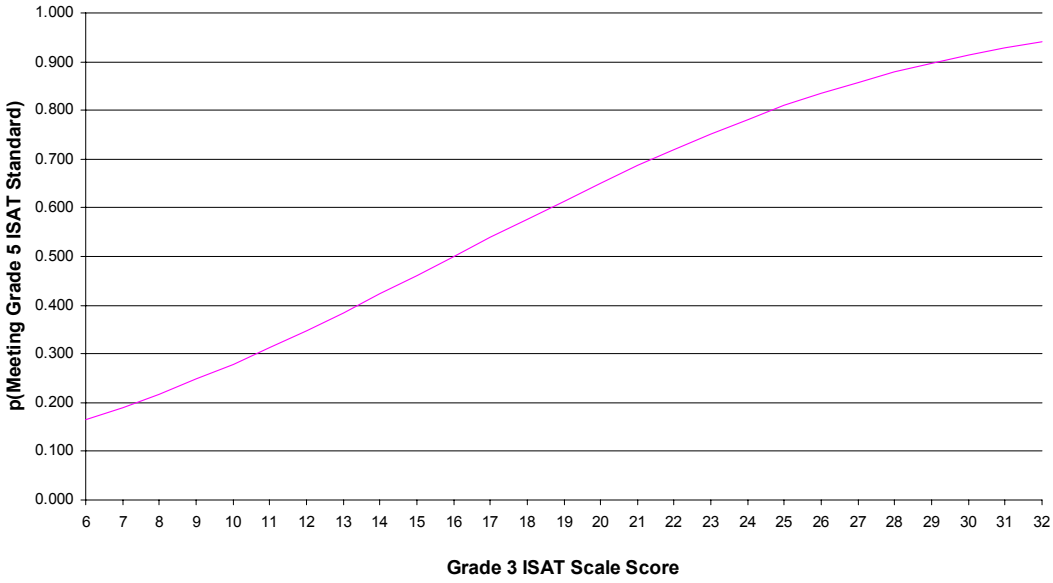


Figure B.3
Relationship Between ISAT Writing Tests



APPENDIX C. IMPACT OF EXTENDED-RESPONSE ITEMS ON READING AND MATHEMATICS SCORE DISTRIBUTIONS

In 2001, extended-response items in the reading and mathematics tests contributed to the overall scale scores for the first time. Two extended-response items have been included in both the reading and mathematics tests since the ISAT program began. However, they were reported separately in the past two years and did not contribute to the overall scale scores.

During this pilot period, a number of studies were conducted to analyze the potential impact of these new items on the reliabilities and distributions of test scores. Table C.1 provides data on the reliabilities of the test scores when the extended-response items are excluded, as they were in the past, and when they are included, as happened in 2001. For reading, there are very slight increases at all three grades when the extended-response items are included. For mathematics, there is a slight decrease for the grade five test. The other grades stay the same. Considering the already high levels of reliability of these ISAT tests, it is perhaps unreasonable to expect any large increase as a result of adding just two more items to the total score. The anomaly at grade five seems to be related to the specific extended-response items used in the test form. A comparison among the difficulty values shown in Tables 3.5 through 3.7 shows that the grade five extended-response items are relatively much less difficult than those at the other two grades.

The last two columns of Table C.1 show person separation values (described earlier in Section 2) for the tests with and without the extended-response items. With the exception of the grade five mathematics test anomaly noted above, these values show even more clearly than the reliability coefficients the increase in precision with the inclusion of the extended-response items.

Table C.1
Comparison of Scale Reliabilities With and Without the Extended-Response Items

	Reliability		Person Separation	
Area/Grade	Multiple-Choice Only	Multiple-Choice + Extended-Response	Multiple-Choice Only	Multiple-Choice + Extended-Response
Mathematics				
3	.93	.93	3.53	3.66
5	.93	.92	3.63	3.36
8	.94	.94	3.87	3.92
Reading				
3	.91	.92	3.23	3.42
5	.92	.93	3.40	3.59
8	.90	.91	3.08	3.21

The extended-response reading items are scored on a holistic, 4-point rubric and, therefore, can contribute a maximum of 8 points to the overall raw score. The extended-response mathematics items are scored on three features: knowledge, strategy, and explanation. Each feature is scored on a 4-point rubric. Therefore, these items can contribute a

maximum of 24 points to the overall raw score. In order to maintain uniformity across subjects and over time in how much the extended-response items contribute to the overall score, a weighting system was adopted to fix their total contribution at 15%.

Table C.2 shows the weights that were applied to the 2001 tests. For the mathematics tests, the number of multiple-choice items, 70, is constant across grades. Consequently, the student's score on the extended-response items was first multiplied by .51 and then added to the multiple choice total. This produced a situation in which the maximum number of points that the extended-response items could contribute was 12.24, which represented 15% of the total number of points (82.24 = 12.24 + 70) in the test. A parallel approach was used with reading. However, since the number of multiple-choice items differs slightly across the three tests, different weights were used with each.

Table C.2
Weighting of Extended-Response Items

Test	Multiple choice raw score points	Extended response raw score points	2001 weight	Norm table upper limit
Reading–3	67	8	1.48	79
Reading–5	66	8	1.46	78
Reading–8	65	8	1.43	76
Mathematics	70	24	0.51	82

The next six charts (Figures C.1-C.6) compare the distributions of scale scores computed with and without the extended-response items. The horizontal axis of each chart represents scale scores in five-point intervals. The point 123, for example, actually represents the score range from 121 to 125. The vertical axis shows the percent of scores falling within each score range category. Two lines are plotted in each chart. The "MC" line shows the distribution of test scores based only on multiple-choice items. The "MC + ER" line shows the distribution of scale scores based on both multiple-choice and extended-response items calculated in accordance with the procedure described above.

As the six charts show, the scale score distributions are quite similar. At the highest score ranges, there is a tendency for the multiple-choice scale scores to produce proportionally more scores than scale scores based on a combination of multiple-choice and extended-response items. However, a review of the item difficulty values reported in Tables 3.2 through 3.7 shows that the extended-response items are typically the most difficult items in each test. The performance of top-scoring students who were not previously challenged is better differentiated when these items are included.

Figure C.1
Comparison of Grade 3 Reading Test Score Distributions

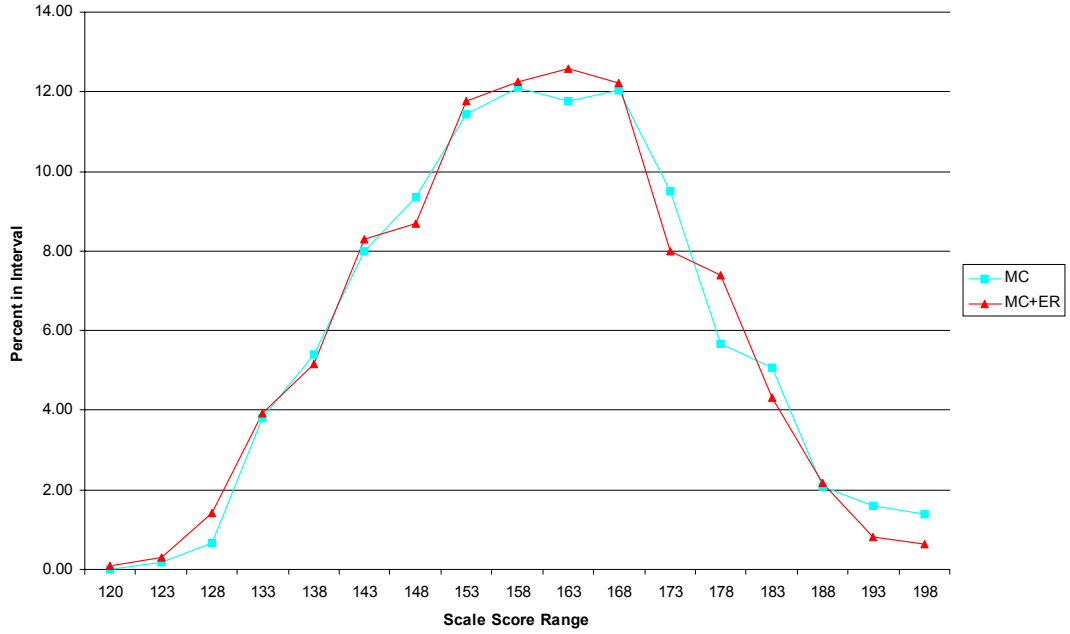


Figure C.2
Comparison of Grade 5 Reading Test Score Distributions

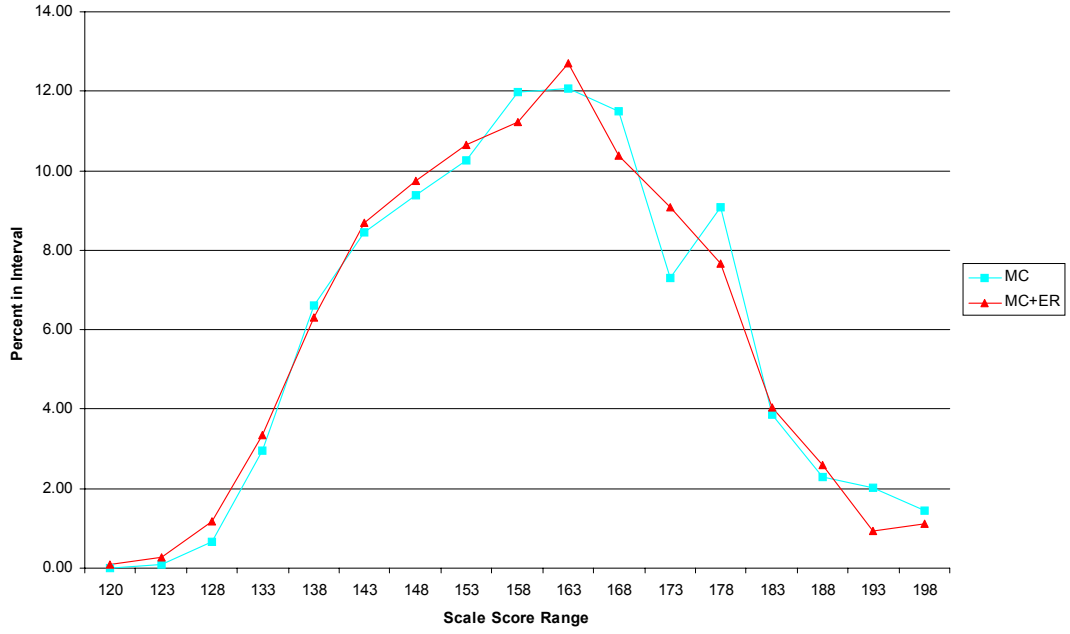


Figure C.3
Comparison of Grade 8 Reading Test Score Distributions

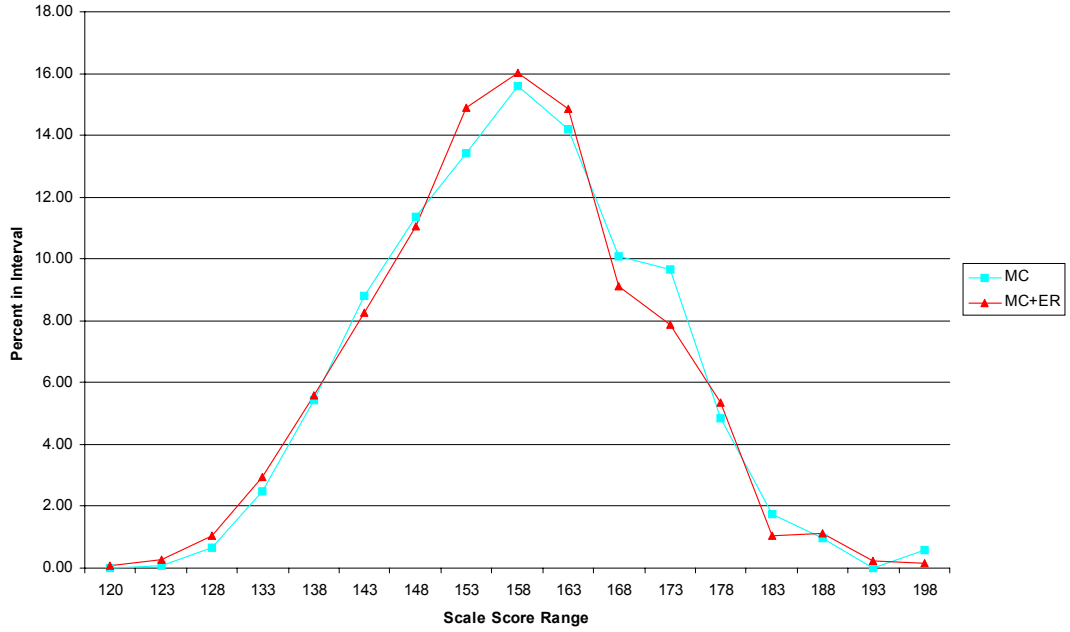


Figure C.4
Comparison of Grade 3 Mathematics Test Score Distributions

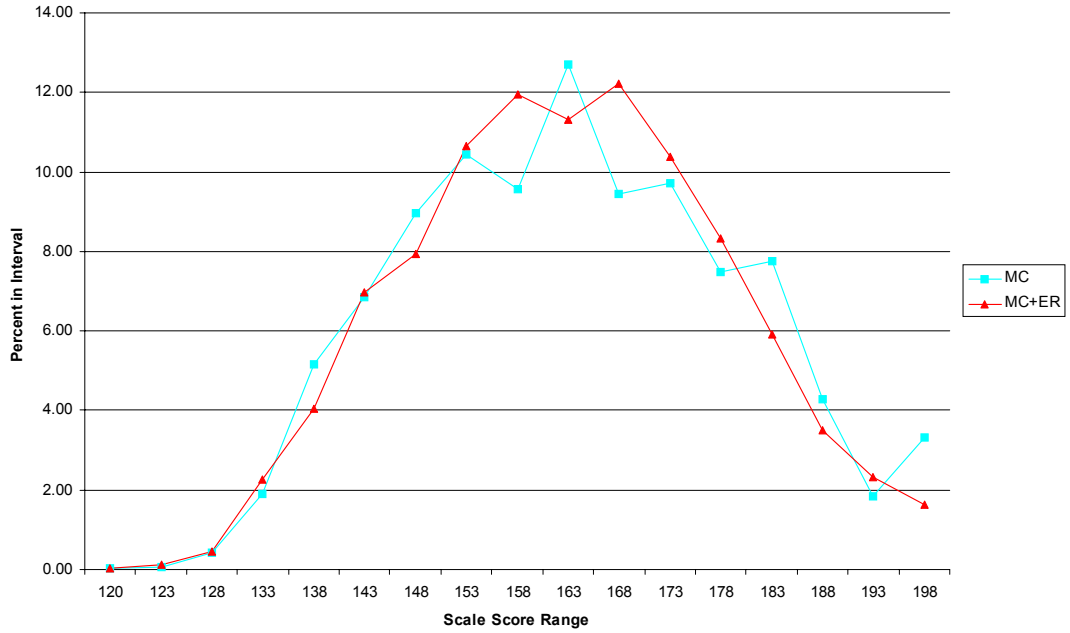


Figure C.5
Comparison of Grade 5 Mathematics Test Score Distributions

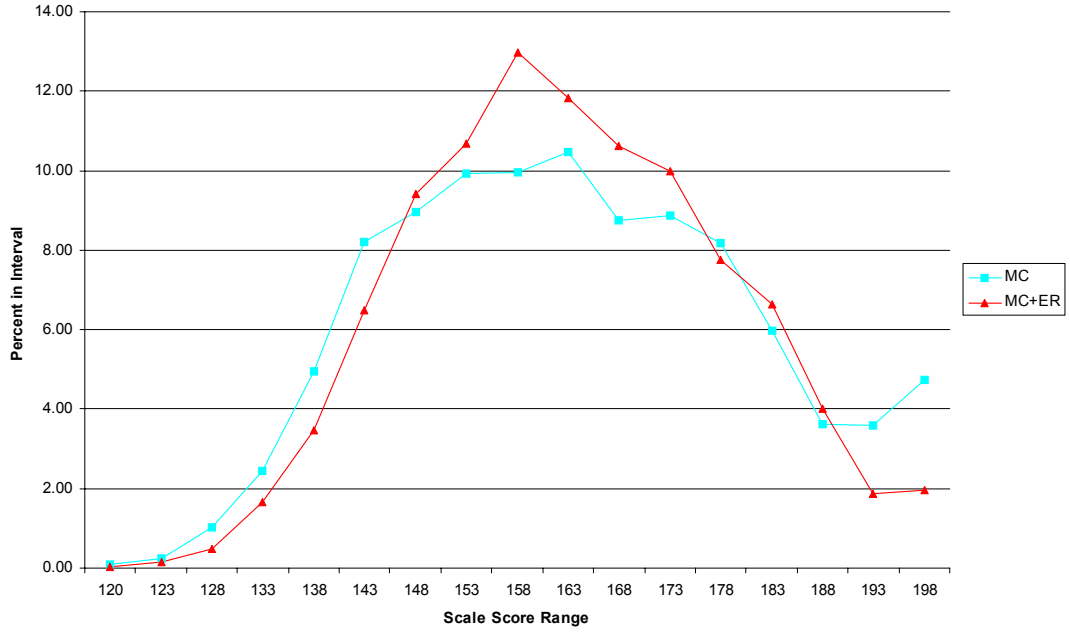


Figure C.6
Comparison of Grade 8 Mathematics Test Score Distributions