Illinois Standards Achievement Test

2014 Technical Manual

Illinois State Board of Education
Division of Assessment

Table of Contents

(* The Appendix C is appended in a separate file.)

1. PURPOSE AND DESIGN OF THE ISAT TESTING PROGRAM

In spring 2014, students in grades 3 through 8 took the Illinois Standards Achievement Tests (ISAT) in reading and mathematics. Students in grades 4 and 7 took the ISAT tests in science as well. Approximately 900,000 students who were 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 (ILS). These standards define what students in all Illinois public schools should know and be able to do in the seven core areas as a result of their elementary and secondary schooling. On June 24, 2010 the Illinois State Board of Education adopted the *Common Core State Standards* for Mathematics and English Language Arts to better prepare students for college and workforce. Prior to these standards, the ILS had not changed since their adoption in 1997. The 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 2014 tests. It describes the tests and assessment approaches and provides evidence of their technical adequacy. Other reports, documents, or publications issued by ISBE provide additional information about how to interpret test results (e.g., *Guide to the 2014 Illinois State Assessment*, *Understanding Your Child's ISAT Scores*), which are not included as part of the report, but can be found from http://www.isbe.net/ils/default.htm

Test Development

Each ISAT test is designed to assess the Illinois Learning Standards validly, reliably, and fairly. The selection of items and assembly of each test is guided by a set of specifications: the Illinois Assessment Frameworks[1]. There are several references to the IL Assessment Frameworks, but the Frameworks are only relevant to science. They do not apply to math or reading.

> The Illinois Assessment Frameworks are designed to assist educators, test developers, policy makers, and the public by clearly defining those elements of the Illinois Learning Standards that are suitable for state testing. They are not designed to replace local curricula and should not be considered state curricula. They define the content that may be assessed on ISAT and PSAE.

These specifications were developed to make certain that test content corresponds to the purposes, objectives, and skills framed by the learning standards (i.e., *Common Core Standards*), and to define those elements of the standards that are suitable for state testing. The census items in the math and reading portions of the 2014 ISAT were totally composed of items written to measure the *Common Core State*

[1] http://www.isbe.net/assessment/IAFIndex.htm

Standards (CCSS). This change was made to speed the transition from the former IAF to a revised framework that reflects the integration of the CCSS into the Illinois Learning Standards.

Illinois teachers and administrators participate in all phases of the test development process: item writing, item selection, bias review, and data review. The State Board of Education convenes a series of advisory committees to build a test development process that 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 State Standards, grade-level appropriateness, importance, and clarity. Incorrect choices (for multiple-choice items) are reviewed for plausibility. In tests other than reading, text complexity of the questions is kept to the minimum necessary to state the problem.

Difficulty. Items are pilot tested on large samples of students to develop a statistical profile for each item prior to their inclusion in tests. 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 often target the use of items with point biserials of at least 0.30 and select items with the highest point-biserials.

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

The 2014 ISAT tests did not contain items from the *Stanford Achievement Test, Tenth Edition* (SAT 10), as in previous test administrations. Instead, the SAT 10 portion of the ISAT tests was replaced with sets of census items from the ISAT item bank that aligned to the Illinois Assessment Framework for science, and the Common Core State Standards for mathematics and English Language Arts. As a result, the 2014 test administration reports students' test performance relative to the Illinois Learning Standards.

ISBE takes several provisions to help ensure test security. Test materials shipped to schools are packaged and sealed. Each test booklet is bar-coded for security purposes. The administration of tests is standardized with a series of manuals providing 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 surplus materials.

Reading

The ISAT reading test assesses material defined by standards associated with two state learning goals: Reading for Literature and Reading for Informational Text. These learning standards are designed to guide language arts instruction in Illinois schools. This alignment of assessment to curriculum ensures consistency and strengthens the influence of standards and assessment on improved teaching and learning. These standards are:

- CCSS RL: Reading Literature
  - Standard: Key Ideas and Details
  - Standard: **C**raft and Structure
  - Standard: Integration of Knowledge and Ideas

- CCSS RL: Reading for Informational Text
  - Standard: Key Ideas and Details
  - Standard: **C**raft and Structure
  - Standard: Integration of Knowledge and Ideas

More detailed information regarding the above reading categories can be found at http://www.isbe.net/common_core/pls/level1/pdf/ela-standards.pdf

The reading test consists of 60 multiple-choice questions and one extended-response question. The census portion of the test is composed of fifty multiple-choice questions and one extended-response question. The remaining ten multiple-choice items are used to link the 2013 ISAT to the 2014 ISAT scales. The test is administered in three 45-minute sessions. Any student who is actively engaged in testing after 45 minutes may be allowed 10 extra minutes to complete that test session.

The reading passages and accompanying questions reflect two of the most frequent purposes for reading—reading to gain information and reading for literary experience. Grade appropriate, high interest passages have been commissioned to be written for exclusive use on the ISAT. A committee of Illinois educators reviewed passage submissions and selected a balance of literary and expository passages for each year's item development.

The multiple-choice questions require students to select one correct response from four options presented to them. Questions must meet both content and statistical criteria for inclusion in the test. The extended-response question on the reading test requires 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 (with 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 the assistance of Illinois educators.

In addition to an overall reading score, results are reported in terms of the percent of items correctly answered within the following categories:

- Reading for Literature and Reading for Informational Text

  Key Ideas and Details

  Craft and Structure

  Integration of Knowledge and Ideas

Mathematics

People use mathematics to identify, describe, and investigate the patterns and challenges of everyday living. 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 observed relationships and problems. Mathematics helps us to understand events that have occurred and to predict and prepare for events to come so that we can understand our world better and live in it more successfully. Illinois mathematics content standards (based on the *Common Core State Standards*) guide teachers to focus their instruction on fewer, higher, and deeper standards than the previous state standards. The ISAT tests measure these key attributes, such as applied problem solving and conceptual understanding. Item formats such as performance tasks and technology-enhanced items to measure standards pertaining to mathematical modeling and long-term research were not used to minimize departures from previous year paper-and-pencil administrations.

In grades 3-5, the ISAT tests measure learning standards for the following domains:

- Operations and Algebraic Thinking
- Number and Operations in Base Ten
- Number and Operations –Fractions
- Measurement and Data
- Geometry

In grades 6-7, the ISAT tests measure learning standards for the following domains:

- Ratios and Proportional Relationships
- The Number System
- Expressions and Equations
- Geometry
- Statistics and Probability

In grade 8, the ISAT tests measure learning standards for the following domains:

- The Number System
- Expressions and Equations
- Functions
- Geometry
- Statistics and Probability

More detailed information regarding the above math domains can be found at http://www.isbe.net/common_core/pls/level1/pdf/math-standards.pdf

The mathematics assessment contains 70 multiple-choice questions, three short-response questions, and two extended-response questions. Five multiple-choice, one short-response, and one extended-response question are pilot-test questions that do not contribute to students' test scores. The test is administered in three 45-minute sessions. Any student who is actively engaged in testing after 45 minutes may be allowed 10 extra minutes to complete that test session. In 2014, the pilot section brought mathematics items to strengthen the linkage of the ISAT scale between the 2013 and the 2014 mathematics test administrations. None of the items placed on the pilot section contributed to students' ISAT mathematics test score.

The multiple-choice questions require students to select one correct response from four options presented to them. Questions must meet both content and statistical criteria for their inclusion in the test. The short-response questions pose similar questions as multiple-choice items but require students to respond without being presented with answer choices. The rubric used to score the short-response items has a scale from 0 to 2 points (with 2 being the highest score). The extended-response questions require students to consider a situation that demands more than a numerical response. The student is required to "solve" the situation, choose a plan, carry out the plan, and interpret the solution derived in terms of the original situation. Students are expected to clearly communicate their decision-making processes in the context of the task proposed by the item. The rubric used to score the extended-response items has three scoring dimensions: Mathematical Knowledge, Strategic Knowledge, and Explanation, with each dimension having a scale from 0 to 4 points (with 4 being the highest raw score). The short-response and extended-response scoring rubrics were developed with the assistance of Illinois educators.

In addition to an overall mathematics score, results are reported in terms of the percent of items correctly answered on each of the following categories:

- Operational and Algebraic Thinking
- Number and Operations in Base Ten
- Number and Operations – Fractions
- Measurement and Data
- Geometry

- Ratios and Proportional Relationships
- The Number System
- Expressions and Equations
- Statistics and Probability
- Functions

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 application of science and in elements of technological design.

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

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

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

- **State Goal 13:** Understand the relationships among science, technology, and society in historical and contemporary contexts.
  Standard 13A: Know and apply the accepted practices of science.

Standard 13B: Know and apply concepts that describe the interaction between science, technology, and society.

The science assessment contains 82 multiple-choice questions; 7 of which are pilot-test questions that do not contribute to students' test scores. The test is administered in two 45-minute sessions. Any student who is actively engaged in testing after 45 minutes may be allowed 10 extra minutes to complete that test session.

In addition to an overall Science score, results are reported in terms of the percent of items correctly answered within five strands. These strands are as follows:

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

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

- *Matter, Energy, and Forces:* Understanding and applying concepts that describe properties of matter and energy and the interactions between them. Knowing and applying concepts that describe force and motion and the principles that explain them. (Standards 12C, 12D)

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

- *Safety, Practice, Science/Technology/Society, and Measurement:* 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, 13B)

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 and each skill applies to both content (knowledge) and to process (research methods). These four skills include: (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 that 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 limited to a lower percentage. 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.

Item Bias Review and DIF Analysis

All ISAT items are screened for potential bias by teacher panels, administrators, and vendor content experts. They are checked during three stages: item writing, item review, and data review. First, all of the teachers who are involved in item writing are trained and instructed to balance ethnic and gender references and to avoid gender and ethnic stereotypes. Then, another group of teachers is invited to the item review meetings to screen for potential language and content bias. Items approved by the item review committee are pilot tested and analyzed for differential item functioning. Last, Illinois administrators, vendor content experts, and a group of Illinois teachers review each item based on statistical inputs in data review meetings.

Differential item functioning (DIF) refers to the different statistical performance of an item between groups of students after differences on test performance have been controlled for the groups. ISAT DIF analyses are done in three ways: males versus females, White versus Black, and White versus Hispanic. The two DIF statistical methods used are Mantel-Haenszel Delta and Mantel chi-square (Angoff, 1993; Dorans & Holland, 1993).

Mantel-Haenszel Delta is used for multiple-choice items. It is transformed from Mantel-Haenszel alpha,

$$\hat{\alpha}_{MH} = \frac{\sum_i p_{ri} q_{fi} N_{ri} N_{fi} / N_i}{\sum_i q_{ri} p_{fi} N_{ri} N_{fi} / N_i},$$

where p_{ri} is the proportion of reference-group students (i.e., male, White) who answered the item correctly in the score-group i, and q_{ri} is 1- p_{ri}. N_{ri} and N_{fi} are the numbers of students in the reference and focal groups, respectively, at each score-group i. Similarly, p_{fi} is the proportion of focal-group students (i.e., female, Black, Hispanic) who answered the item correctly in the score group i, and q_{fi} is 1- p_{fi}. When a constant of -2.35 is applied to the natural logarithm of Mantel-Haenszel alpha, it becomes Mantel-Haenszel Delta (-2.35 $ln[\hat{\alpha}_{MH}]$).

Mantel chi-square is used for open-ended items. Its expression is

$$M-\chi^2=\frac{\left[\left|\sum_m R_{rm}-\sum_m E(R_{rm})\right|-.5\right]^2}{\sum_m Var(R_{rm})},$$

where R_{rm} is the number of reference-group students in score-group m who answered the item correctly, $E(R_{rm})$ is the number of the reference-group students of score-group m expected to answer the item correctly, and $Var(R_{rm})$ is the variance of R_{rm}.

The statistical procedures for DIF analyses are carried out separately for each item in the tests and for several pairs of focal and reference groups. Evaluation of DIF severity involves the use of the well-known ETS DIF categories, A, B, and C, where A represents a negligible DIF, B represents a moderate DIF, and C represents a large DIF (Longford, Holland, & Thayer, 1993).

Table 1.1 and Table 1.2 show data to support the pulling of the 2014 ISAT tests. The data shows numbers of items and quality of the items from the 2013 FT administration available for the pulling of the 2014 tests. Table 1.1 summarizes the number of pilot items that are accepted, rejected, and re-pilot tested. Note that the decisions on pilot items are made based on item *p*-value, point-biserial, and DIF results, not on DIF results alone. None of the rejected items made it to the item bank and test construction. Instead these items were removed from test development activities.

Table 1.1: Data Review Results

Subject	Grade	Total Pilot Items	# Accepted	# Rejected	# Re-Pilot Test
Reading	3	162	132	30	0
	4	160	133	27	0
	5	162	140	22	0
	6	160	131	29	0
	7	162	127	35	0
	8	162	131	31	0
Mathematics	3	70	69	1	0
	4	105	105	0	0
	5	104	104	0	0
	6	108	108	0	0
	7	107	106	1	0
	8	116	115	1	0
Science	4	42	34	8	0
	7	42	28	14	0

Table 1.2 summarizes items selected for operational use in the spring 2014 administration by DIF category B and C using the ETS DIF classification system. The large number of statistical DIF analyses inflates the possibility to unduly flagging items that should not have been flagged, otherwise. To overcome this statistical challenge, panels of teachers and content specialists convene to review data and content of the items showing statistical DIF B and C flags. The purpose of the reviews is to ascertain the presence of content irrelevant sources explaining the DIF flags. Items that survive the content review process become part of the pool of items accessible to test construction activities.

The use of items during test construction is guided with rules available from the psychometric literature (Longford *et al*., 1993). Items from ETS A category are chosen first, and they are the vast majority of the items comprising an operational test. However, when items from the ETS A category are not enough to fulfill test blueprints, items from the ETS B category (that survived the data review process) become candidates for selection. The item with the smallest absolute DIF value is chosen among the competing items. Under extenuating circumstances, with the approval of ISBE few items from the ETS C category (that survived the data review process) become candidates for selection. The use of the item is well documented and its performance is followed across the scoring process. No empirical evidence has been found to support the rejection of items from operational scoring.

Table 1.2: ETS DIF B and C Categories between Male/Female, White/Black, and White/Hispanics

		Male/Female		White/Black		White/Hispanics	
Subject	Grade	B	C	B	C	B	C
Reading	3	0	0	0	0	0	0
	4	2	0	2	0	3	0
	5	3	0	2	0	2	1
	6	2	1	1	0	4	1
	7	3	1	0	1	2	1
	8	6	3	3	1	1	0
Mathematics	3	6	0	14	1	5	0
	4	6	0	13	0	5	0
	5	4	0	16	2	5	0
	6	5	1	5	0	7	0
	7	8	0	9	1	8	0
	8	6	0	3	3	3	1
Science	4	0	0	1	0	0	0
	7	6	0	2	1	4	0

Universal Design and Test Accommodations

The goal of universal design in test development is to maximize accessibility without adaptation or special design. The application of universal design principles offers a test that increases the participation of all students, including those with disabilities and English Language Learners. In practice, universal design considers the needs of different subpopulations to maintain test fairness. A benefit of applying universal design to test development is that the test will better accommodate Braille, audio aids, and visual aids.

The ISAT test development process incorporates the following set of principles and associated guidelines of universal design.

Principle	Guidelines
1. Equitable Use	Provide the same means of use for all users. Avoid segregating or stigmatizing users. Provide equal availability for privacy, security, and safety. Make the design appealing to all.
2. Flexibility in Use	Provide choice in methods of use. Accommodate right- or left-handed access and use. Facilitate the user's accuracy and precision. Provide adaptability to user's pace.
3. Simple and Intuitive	Eliminate unnecessary complexity. Be consistent with user expectations and intuition. Accommodate a range of literacy and language skills. Arrange information in order of importance. Provide effective prompting and feedback.
4. Perceptible Information	Use pictorial, verbal, and/or tactile modes for presentation of essential information. Provide adequate contrast between essential information and its surroundings. Differentiate elements in ways that can be easily described. Provide compatibility with devices used by people with sensory limitations.
5. Tolerance for Effort	Arrange elements to minimize hazards and errors. Provide warnings and fail-safe features. Discourage unconscious action in tasks that require vigilance.
6. Low physical Effort	Allow user to maintain a neutral body position. Use reasonable operating forces. Minimize repetitive actions and sustained physical effort.
7. Size and Space for Approach and Use	Provide a clear line of sight to important elements for any seated or standing user. Make comfortable for any seated or standing user. Accommodate variations in hand and grip size. Provide adequate space for the use of assistive devices or personal assistance.

Source*: Universal Design, Pearson Policy Report (Case, 2003).*

Pearson incorporated these principles and guidelines into item development, production, and administration procedures for the ISAT. The standardized Pearson universal design practice includes: (1) training staff on universal design, (2) screening item content and test booklet layout against universal design guidelines, (3) identifying supplementary materials to accommodate students with special needs, and (4) guarding universal design principles at item review committee meetings.

Pearson's universal design guidelines were implemented in item development for the ISAT by Pearson facilitators. The following considerations are incorporated in the Pearson item development training materials.

1. Considerations for tests
    a. Include and fairly represent as many groups as is reasonable.
    b. Include the numerous perspectives characterized by an issue rather than presenting only one side.
    c. Include a balance of roles for the groups represented. For example, include the contributions of both males and females as well as of various ethnic minority groups.

2. Considerations for items

    Avoid:
    a. descriptions of groups in terms of physical, personality, or interest stereotypes;
    b. the use of language that might be considered derogatory by any group;
    c. the use of words that have different meanings in different cultural settings or dialects;
    d. the use of subject matter likely to be unfamiliar to some groups while familiar to the majority;
    e. the use of esoteric vocabulary or complex sentence structure when that is not being tested; and
    f. the use of material presenting highly controversial or prejudiced points of view.

    Do:
    a. include material relevant to and stressing the positive aspects and values of diversity; and
    b. present positive role models from various groups or material that discusses the contributions of groups to science, history, government, and the arts.

Concepts of universal design are also incorporated in the graphic design of the Illinois test booklet and answer documents, which include:

1. Production
    a. Use a font style that is easy to read.
    b. Enlarge the font size. Note that the previous ISAT font size is similar to the size chosen for the universal design.

c. Design booklet and response sheet to reduce mismatching. Allow large space between items, frame items for easy identification, and use graphic item labels.
d. Choose non-glare paper.
e. Use more dramatic color contrast (including black and white print) to address the needs of different types of color blindness.

2. Administration
   a. Provide adequate testing time.
   b. Repeat instructions.
   c. Incorporate breaks between subtests.

There are five accommodated test formats for special populations: Braille and large print for all subject areas, and reader script, audiocassette, and Linguistically Modified versions of the mathematics and science subtests. Students who take such test formats have additional time as necessary to complete the test. This additional time is determined locally.

Students who take regular test formats have ten minutes of extended time for each test session. The decision of whether to apply the 10-minute extended time period is made at the time of testing by the test administrator, based on whether students are actively engaged in testing after regular time has elapsed.

The Linguistically Modified version of the ISAT was initially introduced in the 2008–2009 school year, and its use has continued through this date. Linguistic modification of test items can be defined as modifying the language of the test to lessen its linguistic complexity while still maintaining the construct of the test. Such modified items avoid linguistic features which increase the reading load of test items, yet have little to do with what the items are supposed to assess. Items were modified (if necessary) using simple, clear, grade-appropriate language and avoiding complex grammatical constructions and idiomatic speech which might be unfamiliar to English language learners. The *ISAT Specifications for Linguistic Modification* was used to train the committee members and guide the process.

ISAT census and pilot test items were reviewed and modified based on language structures/syntax, vocabulary, contextual information, and in some cases formatting to minimize obstacles that may keep students from showing whether they have learned the tested skills.

Language Structures/Syntax

- Test items should be straightforward and easy to understand.
- Use simple and clear language, but avoid choppy sentences.
- Simplify complex sentence structures and avoid compound tenses.
- Use present tense whenever appropriate.
- State the point of the question as early in the sentence as possible.
- Use active voice rather than passive voice whenever possible.
- Limit the use of pronouns. If used, place the pronoun as near as feasible to the referenced noun.

- Avoid contractions.
- Use consistent language structure within an item in order to focus student attention on what is being asked.

Vocabulary

- Use grade-appropriate vocabulary and commonly used words.
- Do not eliminate subject-area terminology that is integral to the skill or concept being assessed.
- When appropriate, use the same word to refer to the same object, phenomenon, etc., throughout the item. Varying words unnecessarily can make text more difficult to understand.
- Avoid using the same word as multiple parts of speech within the same item.
- Avoid words with multiple meanings when their use might be confusing.
- Consider the most commonly understood meaning of a word
- Create and/or label art as needed to help students understand specialized vocabulary that is not content-specific.
- Avoid colloquial and idiomatic language.

Contextual Information

- Avoid using contexts that would be more familiar to some groups of students than to others.
- Delete extraneous information including irrelevant material and unnecessary words in items or graphics.
- Use grade-appropriate, universal contexts that students are likely to encounter in school settings and in textbooks.
- Provide enough contextual information to be clear, but keep in mind that giving too much information can make items lengthy and increase the reading load unnecessarily.

Format

- Determine appropriate font, point size, and use of white space.
- Limit text-wrapping in passages and items.
- Separate text into manageable units (chunking), if needed.

2. RELIABILITY and GENERALIZABILITY

The reliability of a test reflects the degree to which test scores are free from errors of measurement that arise from various sources. Test reliability indicates the extent to which differences in test scores reflect real differences in the construct being measured across some variation in one or more factors, such as time or specific test items used. Different reliability coefficients can be distinguished accordingly. 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. Generalizability theory, which may be thought of as a liberalization of classical theory (Feldt & Brennan, 1989, p. 128), treats these error components and their impact on score precision singly and in interaction.

Internal Consistency of Overall Scores

Because achievement test items typically represent only a relatively small sample from a much larger domain of suitable questions, the test score consistency (generalizability) across items is of particular interest. That is, how precisely will tests line up students if different sets of items from the same domain are used? Unless the lineups 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*, which is quantified in terms of an index called coefficient alpha. The coefficient, which can range from 0.00 to 1.00, corresponds to a generalizability coefficient for a person by item design or, more broadly, as a generalizability coefficient for the person by item by occasion design with one fixed occasion and k randomly selected items (Feldt & Brennan, 1989, p 135). Most well-constructed achievement tests have values above .90. Table 2.1 presents alpha coefficients for the tests administered in the assessment. As the table shows, ISAT tests scores are highly reliable, since the alpha coefficients are comparable to or higher than those typically reported in the literature.

Table 2.1: Reliability Estimates

Grade	Reading	Mathematics	Science
3	0.93	0.93	
4	0.92	0.93	0.92
5	0.91	0.94	
6	0.91	0.93	
7	0.92	0.93	0.93
8	0.92	0.93	

Note: Based on population data

Table 2.1a, Table 2.1b, and Table 2.1c summarize alpha coefficients disaggregated by sub-groups, for ISAT tests by grade. The sizes of the reliability coefficients

remain high for these sub-populations of test takers. In other words, rank-ordering of the test scores remains fairly consistent in each sub-population of students.

Table 2.1a: Reliability Estimates by Ethnicity

Grade	Ethnicity	Reading	Mathematics	Science
3	American Indian or Alaskan Native	0.92	0.92	
	Asian	0.92	0.93	
	Black or African American	0.92	0.92	
	Hispanic	0.91	0.91	
	Native Hawaiian/Pacific Islander	0.92	0.93	
	White	0.92	0.92	
	Two or More Races	0.93	0.93	
4	American Indian or Alaskan Native	0.92	0.92	0.92
	Asian	0.90	0.94	0.91
	Black or African American	0.90	0.91	0.90
	Hispanic	0.91	0.92	0.90
	Native Hawaiian/Pacific Islander	0.89	0.93	0.91
	White	0.91	0.93	0.91
	Two or More Races	0.92	0.94	0.92
5	American Indian or Alaskan Native	0.91	0.93	
	Asian	0.91	0.95	
	Black or African American	0.89	0.91	
	Hispanic	0.90	0.92	
	Native Hawaiian/Pacific Islander	0.88	0.94	
	White	0.91	0.94	
	Two or More Races	0.92	0.94	
6	American Indian or Alaskan Native	0.90	0.91	
	Asian	0.91	0.95	
	Black or African American	0.89	0.90	
	Hispanic	0.90	0.91	
	Native Hawaiian/Pacific Islander	0.90	0.93	
	White	0.90	0.93	
	Two or More Races	0.92	0.94	
7	American Indian or Alaskan Native	0.91	0.92	0.92
	Asian	0.92	0.95	0.92
	Black or African American	0.90	0.90	0.91
	Hispanic	0.91	0.91	0.91
	Native Hawaiian/Pacific Islander	0.93	0.95	0.92
	White	0.91	0.93	0.92
	Two or More Races	0.92	0.94	0.93
8	American Indian or Alaskan Native	0.92	0.93	
	Asian	0.91	0.94	
	Black or African American	0.90	0.91	
	Hispanic	0.90	0.92	

Native Hawaiian/Pacific Islander	0.91	0.93
White	0.91	0.93
Two or More Races	0.92	0.94

Table 2.1b: Reliability Estimates by LEP

Grade	LEP	Reading	Mathematics	Science
3	Yes	0.87	0.90	
	No	0.93	0.93	
4		0.83	0.89	
	Yes	0.91	0.93	0.86
	No	0.78	0.89	0.92
5		0.91	0.94	
	Yes	0.77	0.86	
	No	0.91	0.93	
6		0.79	0.86	
	Yes	0.92	0.93	
	No	0.79	0.87	
7		0.92	0.93	
	Yes	0.87	0.90	0.84
	No	0.93	0.93	0.93
8		0.83	0.89	
	Yes	0.91	0.93	
	No	0.78	0.89	

Table 2.1c: Reliability Estimates by Income

Grade	Low Income	Reading	Mathematics	Science
3	Yes	0.91	0.92	
	No	0.92	0.92	
4	Yes	0.90	0.92	0.90
	No	0.90	0.93	0.90
5	Yes	0.89	0.92	
	No	0.90	0.94	
6	Yes	0.89	0.90	
	No	0.90	0.93	
7	Yes	0.91	0.91	0.91
	No	0.91	0.93	0.91
8	Yes	0.90	0.91	
	No	0.91	0.93	

The reliability coefficients reported in Tables 2.1 to 2.1c are estimated within the context of classical test theory (CTT) and provide single measures of test score precision for the entire test score scale. Within the context of item response theory (IRT), it is possible to measure the relative precision of the test at different points on the test scale. This is often carried out with the test information function, which is a graphical representation of amounts of measurement precision at each test score point (e.g., ability).

Figures 2.1–2.3 show the test information functions for the ISAT reading, mathematics, and science tests. The amount of information at any point is directly related to the precision of the test. That is, precision is the highest where information is the highest. Conversely, where information is the lowest, precision is the lowest, and ability is most poorly estimated. As it is evident from the figures, the information functions for these tests peak near the points on the ability scales where the "Meets Standards" cut scores are located. For example, for reading grade 8, the information function achieves its maximum value near the ability score of -0.01 points which belongs to the "*Meets Standards*" performance level.

Figure 2.1: ISAT Reading Test Information Functions

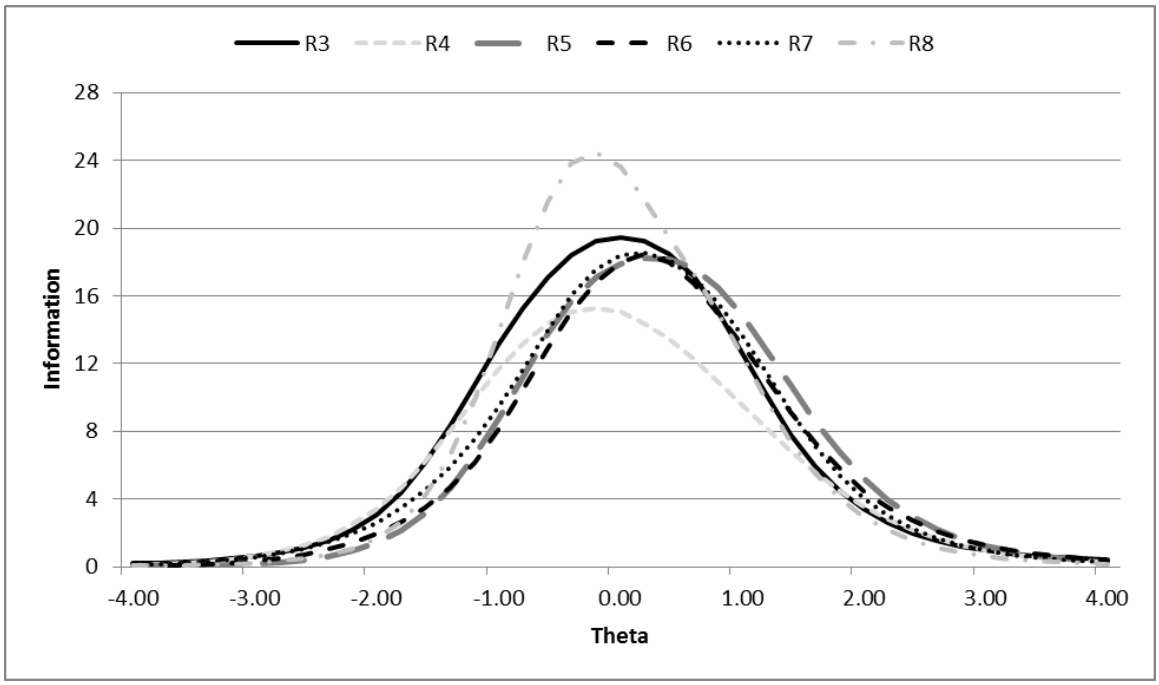


Figure 2.2: ISAT Mathematics Test Information Functions

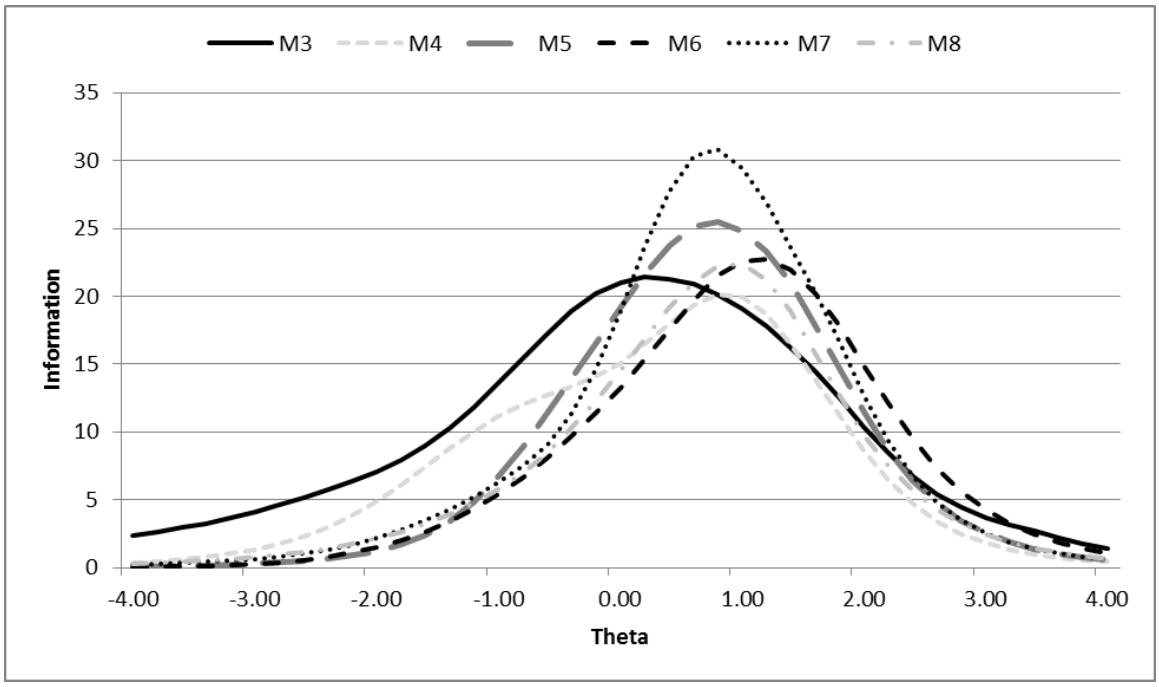
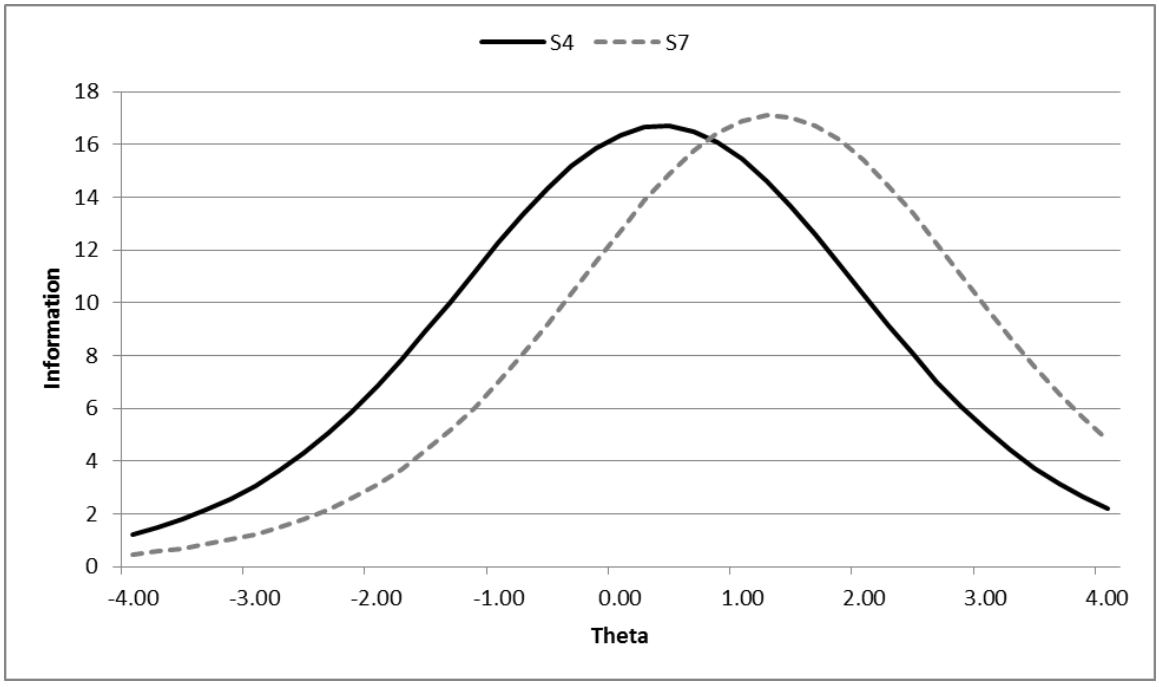


Figure 2.3: ISAT Science Test Information Functions



IRT Conditional SEM

The standard error of measurement (SEM) reflects the degree of error associated with student scores. Classical test theory SEM depicts the amount of measurement error for a typical (average) student disregarding of ability, but item response theory's SEM depicts the amount of measurement error at each point of the ability range. IRT SEM, also known as conditional standard error of measurement (CSEM), is defined as

$$CSEM(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

where $I(\theta)$ is the test information function. The IRT CSEM has an inverse shape relative to the classical test theory's SEM in which for the former SEM values decrease as theta moves toward the center.

Two different procedures are followed to derive the CSEM for ISAT scale scores. The approaches depend on the scaling models underlying the ISAT scale scores. Whereas the ISAT science test is scaled with the Rasch model, the ISAT reading and mathematics tests are scaled with the 3PL/GPC models. For ISAT science, the approach to place CSEM onto the ISAT vertical scale takes the estimates of the CSEM of students' ability and places them onto the vertical ISAT scale by first applying the multiplicative constant for the SAT 10 scale (i.e., 35), and then applying the multiplicative constant for the ISAT scale (i.e., 0.86411). For ISAT reading and mathematics tests, the approach relies on the use of a linear interpolation method and the use of the raw-to-scale table from the previous year. When using this method, the following steps are taken:

1. Obtain raw-to-scale score table for the current year (with proper weighting applied to constructed response items);
2. Map current year's scale scores on previous year's raw-to-scale score table which included the CSEM;
3. If a match is found for a particular scale score, then previous year's CSEM value for that particular score is used;
4. If a match were not found for a particular scale score, then linear interpolation would be used to derive the CSEM for that particular scale score based on the following formula:

   $$CSEM = CSEM_{low} + (CSEM_{high} - CSEM_{low}) \frac{(SCALE - SCALE_{low})}{(SCALE_{high} - SCALE_{low})}$$

   in which SCALE denotes the particular scale score for which the CSEM is to be derived;
   CSEM denotes the CSEM for a particular scale score that is to be derived;
   $SCALE_{low}$ is a scale score from previous year's raw-to-scale score table that is closest on the lower end to the scale score for which the CSEM is to be derived;

$SCALE_{high}$ is a scale score from previous year's raw-to-scale score table that is closest on the higher end to the scale score for which the CSEM is to be derived;
$CSEM_{low}$ is the CSEM associated with $SCALE_{low}$ from previous year's raw-to-scale score table; and
$CSEM_{high}$ is the CSEM associated with $SCALE_{high}$ from previous year's raw-to-scale score table.

The item response theory's SEM is estimated for each reported scale score by subject and grade. The SEM values can be found at Appendix A.

Reliability of the Extended-Response Scores

When test scores are derived from a mixture of multiple-choice and constructed response items, it is important documenting raters' performance and quality of their ratings. Different raters evaluate different students and constructed response items, and when their agreement is low, they can add irrelevance variance to the test scores and thus limit the invariance of students' test scores. Analogously, raters' scores disagreeing with validity scores bring concerns about validity of students' scores. The following section describes the quality control indexes to monitor raters' performance and quality of their ratings.

Inter-rater Agreement

Inter-rater agreement evaluates the consistency of scores assigned to the same response by different readers. For the constructed-response items, inter-rater agreement was monitored daily, and pairs of readers independently scored about 10% of the items' responses across grades.

For the ISAT Reading test, scorers provided a single score for each extended-response item, while for the ISAT Mathematics test extended-response items readers provided three scores: knowledge, strategy, and explanation scores. Tables 2.2 and 2.3 show results for the inter-rater agreement for constructed-response items in reading and mathematics, respectively.

Table 2.2: Inter-rater Agreement for Reading Extended-Response Items

Grade	N	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
3	27722	70	29	99
4	29378	69	30	99
5	29496	67	31	98
6	29668	70	29	99
7	29800	66	33	99
8	29892	68	31	99

Table 2.3: Inter-rater Agreement for Mathematics Constructed- Response Items

Grade	N	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
		Short Constructed Response Item 1		
3	27816	96	4	100
4	29464	96	4	100
5	29510	98	2	100
6	29710	89	11	100
7	29958	92	7	99
8	29886	97	2	99
		Short Constructed Response Item 2		
3	27822	98	2	100
4	29462	92	8	100
5	29520	93	7	100
6	29714	95	5	100
7	29956	94	6	100
8	29980	94	6	100
		Extended Response Item: Knowledge		
3	27710	83	15	98
4	29478	80	19	99
5	29428	90	8	98
6	29714	87	13	100
7	29954	91	8	99
8	30486	89	10	99
		Extended Response Item: Strategy		
3	27710	76	18	94
4	29478	71	25	96
5	29428	86	13	99
6	29714	79	18	97
7	29954	91	9	100
8	30486	88	9	97
		Extended Response Item: Explanation		
3	27710	59	33	92
4	29478	61	33	94
5	29428	62	32	94
6	29714	58	36	94
7	29954	65	26	91
8	30486	63	28	91

The inter-rater agreements on extended-response items are generally found between 91% and 100% for the exact plus adjacent agreements and for the short constructed-response they range between 99% and 100%.

Agreement with Validation Papers

Pearson's validity mechanism provides an objective and systematic check of accuracy. "Validity papers" are actual student responses that are chosen by scoring directors as examples that clearly earn certain scores. These papers are assigned throughout the scoring sessions to monitor raters' accuracy (how raters' scores match validity paper scores). The pool of validity papers includes responses encompassing the entire score range for each item, and scorers read and score them unaware they are scoring validity papers rather than live responses.

Pearson's image scoring system automatically generates a report that compares the scores given by individual scorers with the scores pre-assigned to the validity papers. This report is used to monitor accuracy of individual scorers and the group as a whole. If a scorer drops below an acceptable percentage of accuracy, that scorer may be required to receive individual feedback and/or retraining before being allowed to score any more responses on the given item.

As scoring progresses, additional validity papers are identified through the image scoring system itself. Scoring supervisors use the back-reading tool to identify them to serve as clear examples deserving of certain score points. They regularly escalate such validity papers to scoring directors for their review. Scoring directors select from this pool of validity papers those to be used for validity purposes, choosing valuable examples representing the full range of possible scores. Then, the selected validity papers are transparently routed to all scorers assigned to that item. The validity papers are interspersed with live papers to each scorer at regular intervals throughout the scoring day. Papers in the validity pool are regularly replaced by new samples, which may also be used to target particular scoring issues that arise.

For the ISAT Reading test, scorers provide a single score for the extended-response item, while for the ISAT mathematics test scorers provide three scores for the extended response item: knowledge, strategy, and explanation. Tables 2.4 and 2.5 present agreement with validation papers for extended responses in reading and mathematics, respectively. These values are based on a sample of the total papers scored.

Table 2.4: Agreement with Validation Papers for Reading Extended-Response Items

Grade	N	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
3	6682	81	18	99
4	7152	81	18	99
5	7085	86	14	100
6	6970	89	11	100
7	7384	68	31	99
8	6744	73	26	99

Table 2.5: Agreement with Validation Papers for Mathematics Constructed-Response Items

Grade	N	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
		Short Constructed Response Item 1		
3	1536	99	1	100
4	1629	99	1	100
5	1628	99	1	100
6	1654	97	3	100
7	1669	96	4	100
8	1656	99	1	100
		Short Constructed Response Item 2		
3	1535	99	1	100
4	1632	98	2	100
5	1636	97	3	100
6	1641	96	4	100
7	1666	96	4	100
8	1642	98	2	100
		Extended Response Item: Knowledge		
3	1569	96	4	100
4	1655	96	4	100
5	1648	94	5	99
6	1668	95	5	100
7	1708	96	4	100
8	1717	92	7	99
		Extended Response Item: Strategy		
3	1569	93	6	99
4	1655	94	6	100
5	1648	93	7	100
6	1668	91	8	99
7	1708	96	4	100
8	1717	90	8	98
		Extended Response Item: Explanation		
3	1569	83	14	97
4	1655	76	21	98
5	1648	74	24	98
6	1668	74	22	96
7	1708	79	17	96
8	1717	68	28	96

Reliability of the Performance Category Decisions: Standard Setting

Students' scores on the ISAT tests 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 is briefly described as follows.

Prior to the standard-setting meetings, which took place during April 1999 (for reading and mathematics) and April 2000 (for 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 138 educators participated in the standard-setting process. The distribution of educators across learning areas was as follows: mathematics (56); reading (52); 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 fits the format of the ISAT reading, mathematics, and science 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 should 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—should 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 were originally expressed on the 1999 ISAT scales, which were grade-dependent. With the shift to the 2006 vertical scales, there was a need to conduct a study that would identify points on the new scales that represented comparable levels of achievement. In addition, there was a need to establish corresponding cut points for grades, which were not previously tested (i.e., grades 4, 6, and 7 in reading and mathematics).

The "bridge" study was conducted in 2005. Students who had taken ISAT also completed the SAT 10. The ISAT scores were statistically linked to the SAT 10 vertical scale. Then, when 2006 ISAT results became available, those scores were linked to the SAT 10 vertical scale. This provided the final link to the 2006 ISAT scales, which were linear transformations of the SAT 10 vertical scale. The bridge study results were also used to establish cutoffs for the intermediate grades, which were done by interpolating between existing values.

Results of the bridge study were examined and approved by the State Testing Review Committee at meetings held in September 2005 and January 2006. A panel of content experts also examined these results in December 2005. The State Board of Education voted to accept them at the February 2006 meeting.

In 2013, the ISAT cut scores for reading and mathematics content areas were replaced by a newly adopted set of cuts. These cuts represent higher expectations for Illinois students and they were devised to track students' college and career readiness across the ISAT grade span. Compared to the previous cut scores, the new cut scores raise expectations for the proficient benchmark about 13-17 scale score points in reading and 21-30 scale score points in mathematics. The Illinois Board of Education approved the use of the cut scores on January 23-24, 2013. These cuts remained in place for the 2014 administration. For Science the ISAT cut scores have remained the same as in the previous year.

Table 2.6: ISAT Cut Scores for Each Performance Level

Grade	Academic Warning	Below Standards	Meets Standards	Exceeds Standards
		READING		
3	120-159	160-206	207-235	236-329
4	120-174	175-216	217-248	249-341
5	120-192	193-227	228-260	261-351
6	120-201	202-236	237-266	267-360
7	120-202	203-238	239-270	271-369
8	120-217	218-247	248-270	271-379
		MATHEMATICS		
3	120-172	173-213	214-254	255-341
4	120-190	191-223	224-266	267-355
5	120-200	201-234	235-279	280-369
6	120-213	214-246	247-291	292-379
7	120-220	221-256	257-301	302-392
8	120-233	234-266	267-309	310-410
		SCIENCE		
4	120-157	158-186	187-236	237-361
7	120-196	197-213	214-259	260-390

Source: ISBE document shared to Pearson on January 2013.

The reliabilities of performance level classifications, which are criterion-referenced, are related to the test score reliabilities, but they are not identical. Glaser (1963) was among the first to draw attention to this distinction, and Feldt and Brennan (1989) extensively reviewed the topic.

As Feldt and Brennan (1989, p. 140) point out, approaches to the development of reliability coefficients for criterion-referenced interpretations of test scores have been based either on squared-error loss or threshold loss. It is threshold loss, which evaluates the consistency with which people are 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 coefficients from results of a single test administration (Subkoviak, 1984). An approach proposed by Peng and Subkoviak (1980) was applied to estimate the performance classifications made on the basis of the tests.

Table 2.7 presents the values for p, k, and p_{miss}, the expected proportion of inconsistent decisions, which is simply $(1 - p)$.

Table 2.7: Reliability of Student Performance Decisions Based on 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	0.944	0.598	0.056	0.903	0.804	0.097	0.895	0.663	0.105
	4	0.961	0.599	0.039	0.890	0.777	0.110	0.885	0.584	0.115
	5	0.937	0.582	0.063	0.887	0.768	0.113	0.900	0.588	0.100
	6	0.951	0.581	0.049	0.883	0.761	0.117	0.900	0.578	0.100
	7	0.943	0.578	0.057	0.880	0.751	0.120	0.891	0.542	0.109
	8	0.953	0.658	0.047	0.889	0.774	0.111	0.894	0.558	0.106
Mathematics	3	0.945	0.605	0.055	0.892	0.781	0.108	0.937	0.725	0.063
	4	0.951	0.609	0.049	0.893	0.769	0.107	0.943	0.722	0.057
	5	0.938	0.432	0.062	0.900	0.782	0.100	0.944	0.786	0.056
	6	0.912	0.477	0.088	0.888	0.767	0.112	0.947	0.772	0.053
	7	0.921	0.436	0.079	0.890	0.775	0.110	0.957	0.755	0.043
	8	0.934	0.521	0.066	0.893	0.778	0.107	0.938	0.758	0.062
Science	4	0.972	0.569	0.028	0.908	0.748	0.092	0.909	0.665	0.091
	7	0.953	0.689	0.047	0.922	0.759	0.078	0.900	0.733	0.100
AVERAGE		0.944	0.567	0.056	0.894	0.771	0.106	0.917	0.673	0.083

Note: p and k are estimated for the population of test takers.

In interpreting the first two indices, 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 support consistent classification of students' test performance. For all content grade combinations, the average of the estimates of p ranges from 0.894 to 0.944.

3. VALIDITY

Test validity refers to the degree to which a test measures what it is intended to measure (Cronbach & Meehl, 1955). Evidence that supports a test's validity argument is gathered for different aspects and through different methods. This process is known as validation of test score interpretation and use (Kane, 2013). The two recognized rules to inferring claims on test score use and interpretations are content validity and construct validity. Content validity refers to how well a test covers the content of interest. The process to provide warrants does not involve any statistical computation. Instead, it dwells on logical analyses such as examinations of correspondence between test blueprints and test items. Construct validity is another rule to draw inferences on test scores uses and interpretations. Warrants on construct validity include quantitative analyses such as correlations between items and the test and factor analyses (Cronbach & Meehl, 1955; Crocker & Algina, 1986; and Clark & Watson, 1995).

Content Validity

One piece of evidence on content validity was provided in the form of the 2014 Test Construction Specifications. This document contains descriptions of the blueprint, the process, and the decisions made for defining and developing the ISAT tests. Also, a content validity report was developed to summarize content representation information for the 2014 ISAT *Common Core State Standards*.

Construct Validity

Dimensionality

Dimensionality is a unique aspect of construct validity. Investigation of test dimensionality is necessary for item response theory (IRT) because univariate IRT models assume that a test measures only one latent trait (unidimensionality). Although it is generally agreed that unidimensionality is a matter of degree rather than an absolute condition, there is no consensus on what defines dimensionality or on how to evaluate it. Approaches that evaluate dimensionality can be categorized into answer patterns, principal components, and factor analysis. Principal components and factor analysis are among the most popular methods for evaluation of test dimensionality (Hattie, 1985; Abedi, 1997).

There are alternative rules to evaluate test dimensionality with principal components. Lord (1980) stated that if the ratio of the first to the second eigenvalue is large and the second eigenvalue is close to other eigenvalues, the test is unidimensional. Divgi (1980) expanded Lord's idea and created an index by considering the pattern of the first three factor components (eigenvalues). The Divgi Index examines the ratio of the difference of the first and second eigenvalues over

the difference of the second and third eigenvalues. A large ratio indicates a greater difference between the first and second eigenvalues, thus, creating a unidimensional tendency. A cut value of three is chosen for the index so that values greater than three are indicative of a unidimensional test. Table 3.1 lists results with the Divgi index by grade and subject. All values are greater than 3 supporting that the ISAT tests are essentially comprised by a single dimension. Scree plots, another reference of dimensionality, are presented in Appendix B. The elbow shaped plots support the unidimensionality conclusion drawn from the Divgi index.

Table 3.1: Divgi Index

Grade	Reading	Mathematics	Science
3	29.79	23.86	
4	47.83	14.10	37.37
5	46.49	43.82	
6	50.31	19.91	
7	26.35	35.19	49.90
8	31.62	28.57	

The purpose of studying the internal structure of a test is to demonstrate that all of the items (and groups of items) work coherently. Methods that are used to provide evidence of the internal structure of a test are usually associated with correlations, for example, the item-total correlation and subscale-total Pearson r-correlation.

Empirical data, with all the student population, is used to evaluate test structure through point-biserial correlations of item-total and subscale-total correlations. The subscale scores are the points earned for each reporting category. The corrected point-biserial, in contrast to the uncorrected method, excludes an item from the total score when computing its point-biserial. This method avoids the overestimation issue that commonly occurs in the uncorrected method, and it is carried-out here. The subscale-total correlation includes the subscale items in the total scores.

Table 3.2 shows a summary of item-total point-biserial correlations by grade. The median of the item point-biserial correlations ranges from 0.34 (science) to 0.44 (reading) with a median value of 0.40 across subjects and grades. The minimum value ranges from 0.12 to 0.26. The maximum value ranges from 0.52 to 0.72.

Table 3.2: Median (Min, Max) of Item-Total Point-Biserial by Subject and Grade

	Reading	Mathematics	Science
3	0.44 (0.23, 0.59)	0.38 (0.13, 0.60)	
4	0.41 (0.25, 0.54)	0.39 (0.20, 0.68)	0.34 (0.17, 0.48)
5	0.41 (0.23, 0.52)	0.40 (0.17, 0.72)	
6	0.39 (0.26, 0.56)	0.40 (0.16, 0.62)	
7	0.41 (0.21, 0.55)	0.37 (0.12, 0.71)	0.37 (0.19, 0.53)
8	0.42 (0.20, 0.57)	0.40 (0.23, 0.63)	

(Note: Minimum and Maximum values are shown within parenthesis.)

Table 3.2a shows the mean and standard deviation of item-total point-biserial correlations by grade.

Table 3.2a: Mean (SD) of Item-Total Point-Biserial by Subject and Grade

	Reading	Mathematics	Science
3	0.44 (0.06)	0.39 (0.11)	
4	0.41 (0.07)	0.40 (0.09)	0.35 (0.07)
5	0.40 (0.08)	0.41 (0.11)	
6	0.40 (0.08)	0.39 (0.11)	
7	0.41 (0.06)	0.39 (0.11)	0.37 (0.08)
8	0.42 (0.08)	0.40 (0.08)	

(Note: Standard deviations are shown within parenthesis.)

Tables 3.3 through 3.5 show Pearson r-correlations between test-subscale and total-test scores in reading, science, and mathematics, respectively. The sub-scale scores are reasonably correlated for all content/grade combinations. For example, for reading grade 3 the subscale score correlations range from 0.72 to 0.99.

Table 3.3: Reading Subscale-Total Correlations by Grade

Grade	Subscale	Total	RL	RI
3	Total	1.00	0.97	0.96
	Reading Literature (RL)	0.97	1.00	0.85
	Reading Informational (RI)	0.96	0.85	1.00
4	Total	1.00	0.94	0.96
	Reading Literature (RL)	0.94	1.00	0.81
	Reading Informational (RI)	0.96	0.81	1.00
5	Total	1.00	0.93	0.97
	Reading Literature (RL)	0.93	1.00	0.82
	Reading Informational (RI)	0.97	0.82	1.00
6	Total	1.00	0.90	0.97
	Reading Literature (RL)	0.90	1.00	0.78
	Reading Informational (RI)	0.97	0.78	1.00
7	Total	1.00	0.81	0.99
	Reading Literature (RL)	0.81	1.00	0.72
	Reading Informational (RI)	0.99	0.72	1.00
8	Total	1.00	0.95	0.94
	Reading Literature (RL)	0.95	1.00	0.79
	Reading Informational (RI)	0.94	0.79	1.00

Table 3.5: Science Subscale-Total Correlations by Grade

Grade	Subscale	Total	SI	LE	MF	ES	ST
4	Total	1.00	0.86	0.86	0.83	0.88	0.86
	Scientific Inquiry & Technological Design (SI)	0.86	1.00	0.66	0.63	0.68	0.70
	Life and Environmental Sciences (LE)	0.86	0.66	1.00	0.64	0.70	0.68
	Matter, Energy, & Forces (MF)	0.83	0.63	0.64	1.00	0.67	0.63
	Earth & Space Sciences (ES)	0.88	0.68	0.70	0.67	1.00	0.69
	Safety, Practices, Science/Technology/Society, & Measurement (ST)	0.86	0.70	0.68	0.63	0.69	1.00
7	Total	1.00	0.90	0.85	0.86	0.85	0.89
	Scientific Inquiry & Technological Design	0.90	1.00	0.70	0.72	0.69	0.76
	Life and Environmental Sciences	0.85	0.70	1.00	0.67	0.67	0.70
	Matter, Energy, & Forces	0.86	0.72	0.67	1.00	0.67	0.70
	Earth & Space Sciences	0.85	0.69	0.67	0.67	1.00	0.67
	Safety, Practices, Science/Technology/Society, & Measurement	0.89	0.76	0.70	0.70	0.67	1.00

Table 3.4: Mathematics Subscale-Total Correlations by Grade

Grade	Subscale	Total	OA	NBT	NF	MD	G	RP	NS	EE	SP	F
3	Total	1.00	0.91	0.85	0.79	0.90	0.42	--	--	--	--	--
	Operations & Algebraic Thinking (OA)	0.91	1.00	0.76	0.62	0.70	0.33	--	--	--	--	--
	Number & Operations in Base Ten (NBT)	0.85	0.76	1.00	0.60	0.69	0.31	--	--	--	--	--
	Number & Operations – Fractions (NF)	0.79	0.62	0.60	1.00	0.65	0.33	--	--	--	--	--
	Measurement & Data (MD)	0.90	0.70	0.69	0.65	1.00	0.34	--	--	--	--	--
	Geometry (G)	0.42	0.33	0.31	0.33	0.34	1.00	--	--	--	--	--
4	Total	1.00	0.91	0.82	0.86	0.93	0.63	--	--	--	--	--
	Operations & Algebraic Thinking (OA)	0.91	1.00	0.73	0.72	0.79	0.51	--	--	--	--	--
	Number & Operations in Base Ten (NBT)	0.82	0.73	1.00	0.62	0.71	0.44	--	--	--	--	--
	Number & Operations – Fractions (NF)	0.86	0.72	0.62	1.00	0.72	0.52	--	--	--	--	--
	Measurement & Data (MD)	0.93	0.79	0.71	0.72	1.00	0.53	--	--	--	--	--
	Geometry (G)	0.63	0.51	0.44	0.52	0.53	1.00	--	--	--	--	--
5	Total	1.00	0.73	0.88	0.90	0.91	0.70	--	--	--	--	--
	Operations & Algebraic Thinking (OA)	0.73	1.00	0.62	0.61	0.61	0.48	--	--	--	--	--
	Number & Operations in Base Ten (NBT)	0.88	0.62	1.00	0.71	0.73	0.59	--	--	--	--	--
	Number & Operations – Fractions (NF)	0.90	0.61	0.71	1.00	0.72	0.54	--	--	--	--	--
	Measurement & Data (MD)	0.91	0.61	0.73	0.72	1.00	0.59	--	--	--	--	--
	Geometry (G)	0.70	0.48	0.59	0.54	0.59	1.00	--	--	--	--	--
6	Total	1.00	--	--	--	--	0.82	0.88	0.91	0.92	--	--
	Geometry (G)	0.82	--	--	--	--	1.00	0.64	0.66	0.72	--	--
	Ratios & Proportional Relationships (RP)	0.88	--	--	--	--	0.64	1.00	0.75	0.74	--	--
	Number System (NS)	0.91	--	--	--	--	0.66	0.75	1.00	0.74	--	--
	Expressions & Equations (EE)	0.92	--	--	--	--	0.72	0.74	0.74	1.00	--	--
7	Total	1.00	--	--	--	--	0.76	0.89	0.91	0.90	0.84	--
	Geometry (G)	0.76	--	--	--	--	1.00	0.64	0.60	0.64	0.61	--
	Ratios & Proportional Relationships (RP)	0.89	--	--	--	--	0.64	1.00	0.74	0.73	0.72	--
	Number System (NS)	0.91	--	--	--	--	0.60	0.74	1.00	0.74	0.69	--
	Expressions & Equations (EE)	0.90	--	--	--	--	0.64	0.73	0.74	1.00	0.70	--
	Statistics & Probability (SP)	0.84	--	--	--	--	0.61	0.72	0.69	0.70	1.00	--
8	Total	1.00	--	--	--	--	.88	--	--	.96	.76	.86

Geometry (G)	.88	--	--	--	--	1.00	--	--	.76	.64	.69
Expressions & Equations (EE)	0.96	--	--	--	--	0.76	--	--	1.00	0.66	0.76
Statistics & Probability (SP)	0.76	--	--	--	--	0.64	--	--	0.66	1.00	0.63
Functions (F)	0.86	--	--	--	--	0.69	--	--	0.76	0.63	1.00

4. SCALING AND EQUATING PROCEDURES

Scaling and Equating

ISAT reading, mathematics, and science scores are each reported on a continuous standard score scale. The lowest possible scale score is 120. The upper limit of the scale is restricted to particular values. The restricted scores vary across content/grade combinations but they generally fall below 410 points. The scales are vertically scaled across grades. That is, a score of 235 points, for example, has the same essential meaning for a third-grade student and a fifth-grade student in terms of the achievement it represents.

Because test items and students' ability levels 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 scaling, 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 students' relative performance within a year, but it would not be possible to measure growth across years for students, schools, districts, or the state. The scaling process makes longitudinal comparisons possible.

Starting in 2008, reading and mathematics equating is conducted using the three-parameter logistic model (3-PL model) and the generalized partial credit model (GPC model). Whereas the former allows modeling responses for multiple choice items, the latter allows modeling responses to extended response items. Details of the equating procedure with these two models can be found in the *Documentation of the ISAT Equating for 2008* (Pearson, 2008). The 3-PL model uses item difficulty, item discrimination, pseudo-chance, and the person's proficiency level to describe the probability of a correct response to an item. The GPC model uses all of the above parameters and threshold response category (i.e., difficulty of making a transition from one score point to another) to describe the probability of attaining a particular polytomous item score. Science continues to use the Rasch model. The Rasch model uses only item difficulty and the person's ability to determine the probability of a correct response for a given test item.

The equating procedures may be summarized as follows. Each test form 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, the 3-PL model sets item parameters for these linking items to their historical values through the Stocking-Lord scale transformation coefficients. Test score equating is performed with the true score equating model (Kolen & Brannan, 2004). In the Rasch model the item parameters are set to their historical values through the use of the WINSTEPS constrained calibration approach. By estimating values for the remaining items under these constraints, item parameter values for the remaining items are automatically adjusted to the existing scale. The logic of the equating procedure rests on certain assumptions such as model fit and item

parameter drift. When item parameter drift is present, fluctuations of item difficulty, item discrimination and pseudo-chance have potential to bias the estimated equating function and its precision (Arce-Ferrer & O'Neil, 2012; Wells, Hambleton & Meng, 2011). Careful checks of stability of item parameter estimates are carried out as part of test score equating for reading and mathematics.

Also careful checks are made on the item fit statistics for the anchor items to check data fit to the Rasch model (Arce-Ferrer, 2008). Individual proficiency scores are then transformed using equations developed in the bridge study to have the characteristics of the 2006 reporting scales. The lowest possible scale score is 120, and the student standard deviation of scale scores is approximately 30.

The ISAT has a large testing population and the scoring of their responses is an endeavor that requires time. To decompress the score reporting window, ISAT equating analyses are conducted on students' samples that are drawn from the population of test takers. The sample size is 2,500 students per form when multiple forms are administered and 15,000 students when a single form is administered.

The 2009 and further test administrations are different from previous years in two ways: 1) a linguistically modified form is added to the existing accommodations, and 2) a different cover page is used for accommodations of large print, reader script, and auditory via audiocassette or compact disk (CD) that were not formally distinguished in the 2008 administration. All accommodations that used such a cover are called special form. In 2014, a decision was made to remove the cover for the special form but still track students who took the test under the accommodations. Since 2009, the equating sample is a function not only of the number of ISAT regular forms but also a function of the linguistically modified form, and the special form. As a result of the change introduced in 2014, students taking the ISAT with the same accommodations as those used for the special form became part of the sample. Since 2009, the n- counts for the regular form are still targeted at 2,500 per form (or 15,000 with a single form administration). Samples of the linguistically modified form and the special form (or groups of students taking the ISAT with same accommodations as for the special form) are drawn to reflect their population proportion relative to the regular form proportion. The total sample sizes range approximately from 16,000 to 18,000 students for each grade.

Table 4.1 shows the summaries of the scaled item parameters for reading and mathematics and the Rasch equating results for science from the 2013 operational administration. The item count (*N*), minimum value (*Min*), maximum value (*Max*), mean, and standard deviation (SD) are presented for each of the three parameters. The item discrimination parameter provides information about how well an item discriminates among individuals located at different points. The discrimination parameter can theoretically take values within $-\infty$ and $+\infty$. Similar to classical test theory item discrimination, an item with a negative discrimination parameter should not be used in the test. As positive values of discrimination increases, the item ability to differentiate students at different ability locations increases.

The item difficulty parameter describes the location of the item along the ability continuum and it is helpful to describe the probability of answering correctly an

item. The theoretical range of the item difficulty parameter is from -∞ to +∞ with the positive region indicating difficult items and the negative region indicating easy items.

The pseudo-chance item parameter describes the probability of getting an item correct by pure random chance. Sometimes, this item property is used to describe the probability for an extremely low performing test taker to correctly answer an item. This item parameter can take values between 0.0 and 1.0, with larger values indicating examinees' greater amounts of guessing behavior.

The Rasch model assumes items are equally discriminating with no guessing taking place, and the science portion of the table leaves blank the portion on discrimination and pseudo-chance to keep consistency.

Table 4.1: Summary of Equating Results BY Subject and Grade

			Item Discrimination (a)				Item Difficulty (b)				Pseudo-Chance (c)			
Subject	**Grade**	***N***	***Min***	***Max***	***Mean***	***SD***	***Min***	***Max***	***Mean***	***SD***	***Min***	***Max***	***Mean***	***SD***
R	3	51	0.56	1.96	1.00	0.32	-1.96	1.04	-0.37	0.71	0.00	0.34	0.18	0.09
R	4	51	0.42	1.53	0.86	0.25	-2.07	1.21	-0.44	0.76	0.00	0.46	0.17	0.09
R	5	51	0.37	1.56	0.95	0.26	-1.18	1.41	-0.05	0.66	0.00	0.57	0.19	0.10
R	6	51	0.41	1.80	0.94	0.32	-1.62	1.04	-0.16	0.63	0.00	0.42	0.18	0.10
R	7	51	0.31	1.92	0.95	0.32	-1.92	0.93	-0.35	0.76	0.00	0.31	0.15	0.09
R	8	51	0.38	1.69	1.04	0.30	-2.42	1.15	-0.41	0.70	0.00	0.41	0.18	0.08
M	3	70	0.29	1.71	0.84	0.30	-4.37	2.11	-0.35	1.15	0.00	0.50	0.15	0.10
M	4	70	0.47	1.81	0.91	0.30	-2.71	1.50	-0.18	0.99	0.00	0.43	0.18	0.10
M	5	70	0.38	1.69	0.96	0.30	-2.99	1.69	0.27	0.80	0.00	0.37	0.18	0.09
M	6	70	0.41	1.70	0.97	0.31	-2.31	2.33	0.49	1.01	0.00	0.40	0.20	0.09
M	7	70	0.30	1.84	1.09	0.33	-1.78	2.21	0.25	0.92	0.00	0.54	0.19	0.10
M	8	70	0.41	1.84	0.87	0.29	-3.04	1.95	0.14	0.91	0.00	0.47	0.21	0.11
S	4	75					-1.45	1.81	0.22	0.75				
S	7	75					-0.86	2.77	1.18	0.74				

(Note: Mathematics ER items are scored in three different domains.)

Prevention and Detection of Scale Drift

Scale or item parameter drift is used to describe a condition under which scale scores or cutoff levels on a test do not represent comparable levels of proficiency at two points in time. Under conditions of scale drift, if average scores increase (or decrease) or the proportion of the population scoring above certain target levels changes over time, there can be no confidence that the change represents a real change in knowledge of the material being tested.

There are many valid reasons why scores increase over time, such as improved mastery of the concepts and knowledge represented by the test blueprint and better test preparation. However, the situation may also occur for unacceptable reasons. The scaling of successive test forms, for example, always entails some degree of statistical error, which may accumulate undesirably over periods of time. The frequent repetition of items can also lead to situations where score increases reflect

familiarity with specific content rather than greater familiarity with the underlying subject matter.

The ISAT program takes a number of steps to attempt to reduce the effects of scale drift. The items used to link each successive form represent the range of content being tested and occupy the same positions in different test forms to avoid parameter shifts arising from location differences. The anchor item set is always large, usually with length of at least one-quarter of the test length. During the calibration runs, item parameter stability is carefully and systematically examined to identify any items that appear to have changed in performance. All of these procedures help to safeguard against the undesirable effects of scale drift.

Evaluating a Vertical Scale

Three properties are used to evaluate a vertical scale: grade-to-grade growth, grade-to-grade variability, and the effect size for grade-to-grade differences (Kolen & Brennan, 2004). The grade-to-grade growth and variability of each ISAT test are presented in Figures 4.2 and 4.3 below. The growth is indicated by using the grade level mean scale score and a variability of one standard deviation. Although statistics for ISAT science are included in this session, discussions of these statistics are excluded because the gap exists between grades 4 and 7.

Yen (1986) proposed an effect size index to detect the separation of grade distributions. The effect size computation utilizes the mean, variance, and sample size

$$effectsize = \frac{\bar{x}_{upper} - \bar{x}_{lower}}{\sqrt{(n_{upper}s^2_{upper} + n_{lower}s^2_{lower})/(n_{upper} + n_{lower})}},$$

where x, s^2, and n are the mean, variance, and sample size of the upper and lower grades. This index gives effect size in standard deviation units. Cohen (1988) suggested that the cuts for small, medium, and large effect sizes are 0.2, 0.5, and 0.8, respectively.

Table 4.2 presents the means and standard deviations for each grade and Table 4.3 shows the effect size of grade-to-grade differences. Reading and mathematics show larger rates of growth in the lower grades, but the rates slowdown in the higher grades for reading and some for some grades in mathematics. All of the effect sizes of reading and mathematics are smaller than 1 but greater than 0.2. In other words, the most growth for reading and mathematics is more than 0.2 but less than 1 standard deviation. Based on Cohen's rules, the magnitude of growth can be understood to be between small and medium sizes. The effect size values are consistent to previous year findings and with values reported by Downing and Haladyna (2006).

Table 4.2: Scale Score Means and Standard Deviations by Subject and Grade

	Reading			Mathematics			Science		
Grade	N	Mean	SD	N	Mean	SD	N	Mean	SD
3	147548	207.06	30.61	148318	217.83	31.45			
4	148526	219.12	27.95	149274	232.52	28.82	149193	208.23	29.12
5	147421	231.01	27.00	147847	247.04	31.88			
6	147873	239.64	24.18	148597	254.70	32.25			
7	149441	242.55	26.17	149917	262.86	29.52	149917	239.18	31.37
8	149479	249.34	22.05	150201	276.25	32.56			

Table 4.3: Effect Size of Grade-to-Grade Difference

Grades	Reading	Mathematics	Grades	Science
3-4	0.41	0.49	4&7	1.02
4-5	0.43	0.48		
5-6	0.34	0.24		
6-7	0.12	0.26		
7-8	0.28	0.43		

Figure 4.1: Reading Scale Score Mean and 1-SD Band across Grades

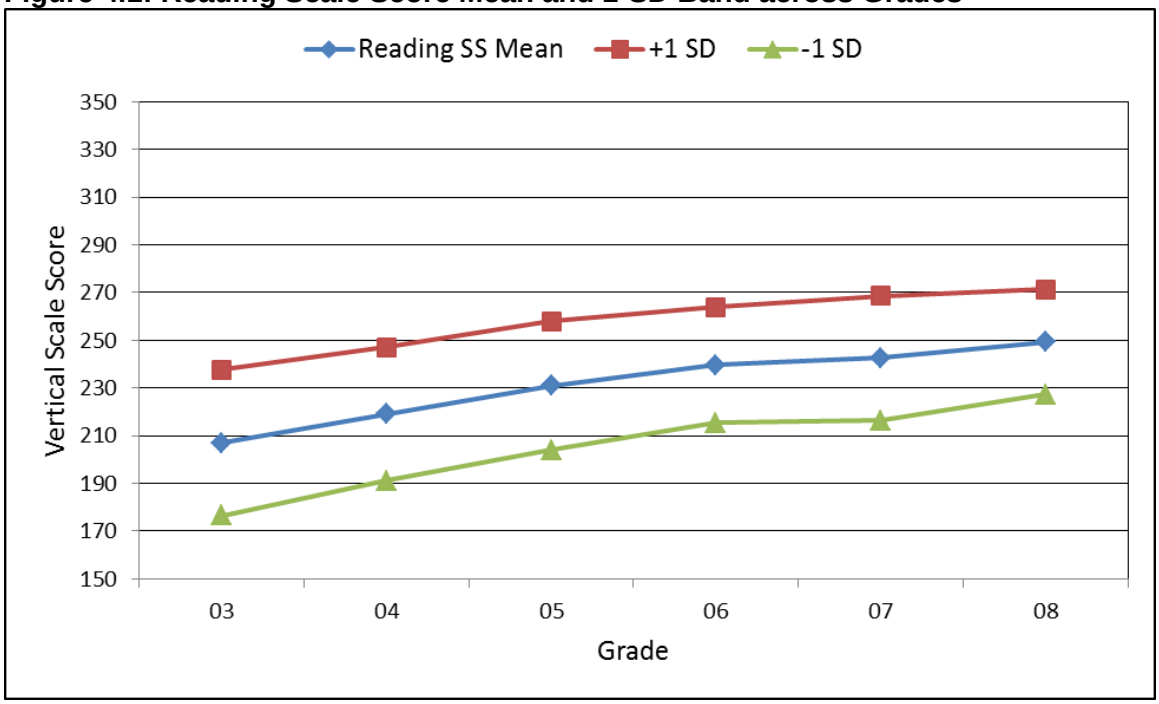


Figure 4.2: Mathematics Scale Score Mean and 1-SD Band across Grades

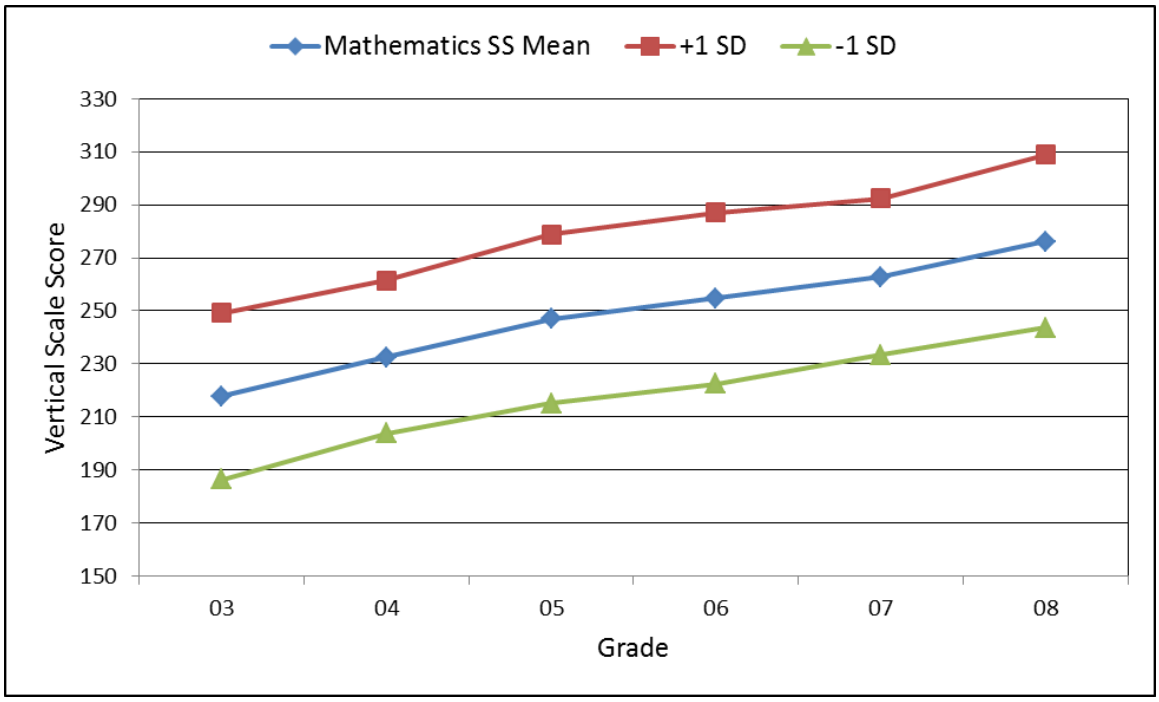
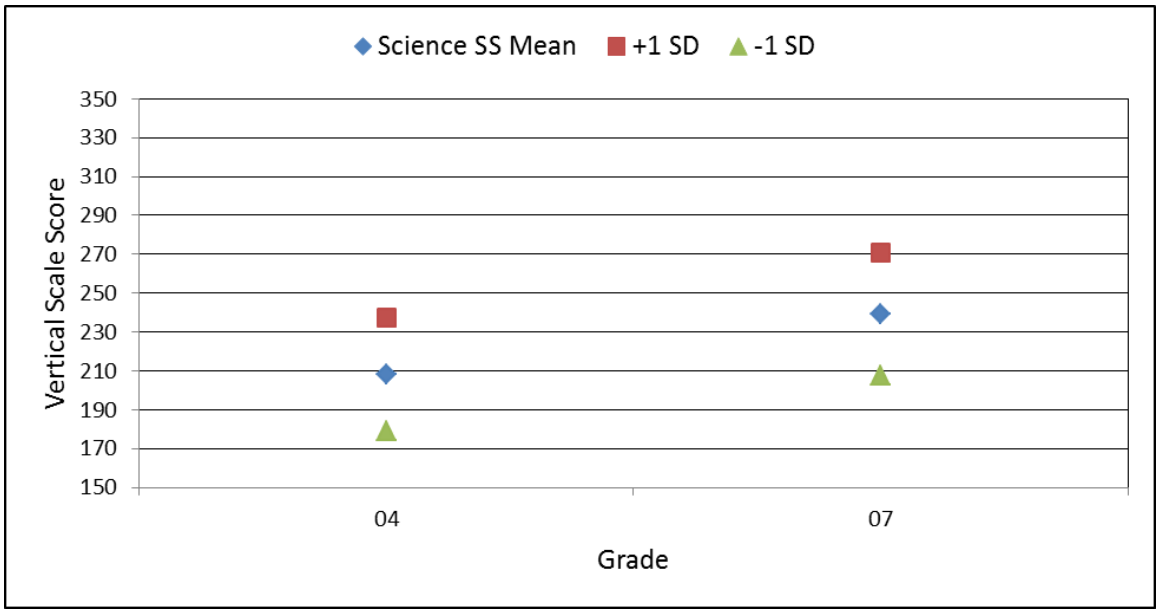


Figure 4.3: Science Scale Score Mean and 1-SD Band across Grades



5. RESULTS

Performance Relative to the Illinois Learning Standards

The longitudinal performance of Illinois students relative to the Illinois Learning Standards can be better understood when considering the history of improvement changes made to the state assessment program. First, due to the cancellation of the IMAGE test, those English language learner (ELL) students who would take the IMAGE started to take the ISAT tests in 2008. Second, beginning 2009, linguistically modified forms of the ISAT tests were administered to the ELL population in mathematics and science. Students who take the linguistically modified forms were included in the operational equating along students taking the regular and special forms. Third, in 2013, the ISAT cut scores for reading and mathematics content areas were replaced by a newly adopted set of cuts approved by ISBE. These cuts pose high achievement expectations to Illinois students. For Science the ISAT cut scores remain the same as in previous year. Fourth, in 2014, the ISAT Reading and Mathematics tests measure Illinois Learning Standards aligned to Common Core State Standards (CCSS). The tests were built with Illinois Assessment Frameworks develop to measure CCSS.

Table 5.1 shows longitudinal track of percentages of students falling into each performance level by subject and grade from 1999, when the ISAT started, through this most recent administration. In order to highlight the change in population, years 1999 through 2007 are shaded to indicate populations before the change from the IMAGE test to the linguistically modified ISAT test. Also, the change in the cut scores beginning 2013 and the use of Illinois Assessment Frameworks aligned to CCSS that took place in 2014 is highlighted to separate those administrations from the previous administrations.

Table 5.1: Percentages of Students by Subject and Grade Falling into Each Performance Level: 1999-2014

		Reading				Mathematics				Science			
Grade	Year	Warning	Below	Meet	Exceed	Warning	Below	Meet	Exceed	Warning	Below	Meet	Exceed
3	1999	8	31	44	17	12	20	47	21				
	2000	6	32	41	21	10	21	46	23				
	2001	7	31	43	19	8	18	46	28				
	2002	7	31	44	19	7	19	44	30				
	2003	8	30	40	22	7	17	45	31				
	2004	7	28	42	23	7	14	46	33				
	2005	7	27	45	22	5	15	45	34				
	2006	6	24	47	23	4	11	47	38				
	2007	5	22	49	24	4	10	45	42				
	2008	7	22	48	24	3	11	44	41				

		Reading				Mathematics				Science			
Grade	Year	Warning	Below	Meet	Exceed	Warning	Below	Meet	Exceed	Warning	Below	Meet	Exceed
	2009	5	23	46	26	3	11	44	41				
	2010	5	21	46	28	3	11	45	42				
	2011	6	19	48	27	3	10	43	44				
	2012	5	19	46	30	3	9	45	42				
	2013	7	34	39	19	7	38	44	11				
	2014[1]	7	37	36	19	7	37	42	13				
4	2000									1	35	51	13
	2001									8	26	54	11
	2002									8	25	53	14
	2003									7	27	52	14
	2004									6	26	55	13
	2005									5	24	55	16
	2006	2	26	47	26	2	14	59	26	3	17	64	15
	2007	1	25	48	25	1	12	57	29	4	17	62	18
	2008	2	25	47	27	1	15	58	26	4	20	59	17
	2009	1	25	46	28	1	13	58	28	3	20	59	18
	2010	1	25	45	29	1	13	58	28	3	20	60	17
	2011	1	24	45	30	1	11	60	28	3	17	58	21
	2012	1	23	47	29	1	11	57	31	3	17	60	20
	2013	6	35	44	15	6	33	48	12	2	17	60	21
	2014[1]	5	39	40	17	7	30	52	12	3	20	60	16
5	1999	1	38	37	24	6	39	53	3				
	2000	0	41	39	20	6	37	52	5				
	2001	1	40	34	25	4	34	55	6				
	2002	1	39	37	22	5	32	55	8				
	2003	1	39	37	23	4	28	59	10				
	2004	2	37	36	25	3	25	60	12				
	2005	2	35	43	19	3	24	61	12				
	2006	1	30	46	22	1	21	64	15				
	2007	1	30	44	26	1	17	63	20				
	2008	1	26	46	27	1	18	64	17				
	2009	0	26	48	26	0	17	66	16				
	2010	0	25	45	30	0	16	66	18				
	2011	0	23	49	27	1	15	65	19				
	2012	0	22	47	31	1	16	66	18				
	2013	6	35	43	16	7	33	48	12				
	2014[1]	8	34	44	14	6	30	49	15				
6	2006	0	27	53	19	1	20	63	16				
	2007	0	26	54	19	1	18	62	19				
	2008	0	21	53	26	1	17	62	21				
	2009	0	20	53	27	1	17	59	23				
	2010	0	19	55	26	1	15	60	24				
	2011	0	16	57	27	1	15	58	26				
	2012	0	18	57	25	0	15	59	26				
	2013	6	35	43	16	7	33	47	13				
	2014[1]	6	37	43	14	9	31	47	14				

		Reading				Mathematics				Science			
Grade	Year	Warning	Below	Meet	Exceed	Warning	Below	Meet	Exceed	Warning	Below	Meet	Exceed
7	2000									12	16	54	18
	2001									11	17	52	20
	2002									10	17	56	17
	2003									10	17	56	18
	2004									10	15	58	17
	2005									10	15	54	20
	2006	1	28	60	12	3	21	55	21	6	13	62	19
	2007	1	26	58	15	2	18	54	25	7	14	55	24
	2008	1	22	59	19	2	18	54	26	6	14	56	23
	2009	0	22	57	21	2	16	55	28	7	14	56	24
	2010	0	22	58	20	2	14	56	28	5	12	60	22
	2011	0	21	58	21	2	13	54	30	6	12	58	24
	2012	0	21	58	20	1	14	54	31	9	12	55	25
	2013	7	35	44	15	7	34	47	12	7	14	54	25
	2014[1]	7	33	45	14	7	35	48	10	8	12	55	25
8	1999	1	27	54	18	5	52	36	7				
	2000	0	28	56	16	8	46	35	12				
	2001	1	34	56	10	7	42	37	13				
	2002	1	31	58	10	7	40	37	15				
	2003	1	36	54	10	6	41	38	16				
	2004	2	31	57	10	6	40	38	17				
	2005	1	27	61	12	6	40	37	17				
	2006	0	21	70	9	2	20	53	26				
	2007	1	18	70	12	1	18	52	29				
	2008	0	18	73	8	2	18	53	27				
	2009	0	16	75	9	1	18	55	27				
	2010	0	15	72	12	1	16	53	31				
	2011	0	15	75	10	0	13	55	32				
	2012	0	14	76	10	0	15	52	33				
	2013	6	34	42	18	5	36	46	13				
	2014[1]	7	36	42	14	7	33	45	15				

Note: 1. The percentages reported in this table reflect the status from data as July 25, 2014. Any corrections to these data after this date might change the percentages in the table.

Table 5.2 presents the average proportion correct of multiple-choice items by reporting categories for the population of Illinois students who took the ISAT test in spring 2014. The proportion correct of a reporting category is the score earned in the category divided by its maximum possible score.

The reporting categories for reading are: 1. Reading Literature and 2. Reading Informational. The reporting categories for mathematics are: 1. Operations and Algebraic Thinking, 2. Number and Operations in Base Ten, 3. Number and Operations –Fractions, 4. Measurement and Data, 5. Geometry, 6. Ratios and Proportional Relationships, 7. Number System, 8. Expressions & Equations, 9. Statistics and Probability, and 10. Functions.

The reporting categories for science include 1. Scientific Inquiry and Technological Design, 2. Life and Environmental Sciences, 3. Matter, Energy, and Forces, 4. Earth and Space Sciences, and 5. Safety, Practice, Science/Technology/Society, and Measurement.

Table 5.2: Average Proportion Correct by Reporting Category

Subject	Reporting Category	Grade					
		3	4	5	6	7	8
Reading	1. Reading Literature	0.65	0.73	0.67	0.71	0.78	0.75
	2. Reading Informational	0.67	0.61	0.60	0.62	0.63	0.60
Mathematics	1. Operations and Algebraic Thinking	0.70	0.65	0.59	0.65	0.57	0.68
	2. Number and Operations in Base Ten	0.69	0.76	0.60	0.67	0.55	0.55
	3. Number and Operations – Fractions	0.47	0.54	0.50	0.58	0.59	0.53
	4. Measurement and Data	0.48	0.58	0.51	0.64	0.58	0.63
	5. Geometry	0.62	0.57	0.64	0.44	0.47	0.54
	6. Ratios and Proportional Relationships				0.58	0.53	
	7. The Number System				0.57	0.65	
	8. Expressions and Equations				0.48	0.50	0.54
	9. Statistics and Probability					0.50	0.63
	10. Functions						0.61
Science	1. Scientific Inquiry and Technological Design		0.71			0.64	
	2. Life and Environmental Sciences		0.65			0.69	
	3. Matter, Energy, and Forces		0.59			0.63	
	4. Earth and Space Sciences		0.62			0.66	
	5. Safety, Practice, Science/Technology/Society, and Measurement		0.71			0.70	

Performance Relative to National Quarters

The legislation that authorized the development of the 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 concurrently with the ISAT to a sample of Illinois students until after 2005. 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.

Between the years 1999 through 2005, the ISAT used the *Stanford Achievement Test*, *Ninth Edition* (SAT 9) for the purpose of determining Illinois students' relative standing within the national population. Equipercentile methodology was used to connect scores on the two tests. In equipercentile linking, the scores on two tests are assumed to be equivalent if they have the same percentile rank. For example, the SAT 9 score that cuts off 10% of the sample is assumed to represent a level of proficiency equal to the ISAT score that cuts off 10% of the sample, even though the scores themselves may be quite different numerically.

Starting in 2006 and ending in 2013, the *Stanford Achievement Test*, *Tenth Edition* (SAT 10) is embedded in the ISAT to provide both criterion- and norm-referenced scores. The SAT 10 national norm is computed solely based on SAT 10 items. Consequently, students of the same ISAT scale scores might receive different national norm scores. Longitudinal track of national quarters of SAT 10 outcomes are shown in Tables 5.3. Since ELL students take regular ISAT reading test and receive linguistically modified mathematics and science ISAT tests, the SAT 10 national quarter for reading includes the ELL population while mathematics and science excludes it. Table 5.3 shows shaded values for the interval 2008 to 2013 for reading. The SAT 10 was not administered in 2014 and Table 5.3 reflects that decision.

Table 5.3: Percentages of Students Falling into Each National Quarter: 1999-2013

		Reading				Mathematics				Science			
Grade	Year	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
3	1999	22	22	25	32	19	21	28	32				
	2000	21	21	25	33	18	21	26	36				
	2001	21	22	25	32	14	19	25	42				
	2002	21	21	26	33	13	19	25	43				
	2003	22	20	25	33	12	18	25	44				
	2004	19	20	26	35	10	17	28	46				
	2005	18	21	23	37	9	18	27	47				
	2006	12	20	32	35	16	18	27	39				
	2007	12	21	33	35	16	19	27	38				
	2008	10	21	30	40	13	18	25	44				
	2009	9	21	29	41	16	15	25	44				
	2010	8	20	29	43	15	15	25	45				
	2011	8	21	25	46	14	15	26	44				
	2012	8	19	26	48	14	15	25	46				
	2013	8	19	25	48	15	16	25	45				
4	2000									18	26	25	31
	2001									19	23	27	30
	2002									18	24	27	30
	2003									18	25	25	32
	2004									16	26	26	32
	2005									13	25	25	37
	2006	9	18	31	43	10	17	32	42	12	23	28	37
	2007	9	17	31	43	10	16	31	43	11	22	29	39

		Reading				Mathematics				Science			
Grade	Year	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	2008	8	19	27	45	10	17	28	44	9	22	28	41
	2009	8	19	27	46	11	19	24	45	13	18	35	34
	2010	7	18	28	47	10	17	29	44	12	18	36	34
	2011	6	15	29	51	6	17	32	45	10	23	28	40
	2012	6	14	33	47	4	17	25	53	8	20	30	42
	2013	6	14	33	47	5	18	25	52	8	20	31	41
5	1999	21	23	27	28	20	22	24	33				
	2000	21	26	28	25	19	22	21	38				
	2001	25	21	24	30	17	19	21	42				
	2002	23	23	26	28	16	19	22	43				
	2003	23	22	27	28	13	17	21	49				
	2004	22	23	27	28	10	16	24	49				
	2005	21	22	33	24	11	15	22	53				
	2006	13	18	33	37	14	16	25	45				
	2007	12	17	33	38	12	15	25	48				
	2008	9	17	27	47	10	15	25	50				
	2009	9	17	27	47	11	17	26	47				
	2010	8	16	27	48	9	15	25	51				
	2011	6	17	26	51	6	18	28	48				
	2012	5	14	25	56	12	15	23	50				
	2013	5	14	25	55	12	15	24	49				
6	2006	13	26	36	24	15	18	30	36				
	2007	13	26	37	25	14	18	30	38				
	2008	9	21	38	33	9	17	24	49				
	2009	9	20	38	33	9	17	24	50				
	2010	8	19	38	35	8	16	24	52				
	2011	7	20	37	37	7	15	24	54				
	2012	6	22	33	40	7	15	28	51				
	2013	6	21	33	40	7	15	28	50				
7	2000									14	24	22	41
	2001									12	25	20	43
	2002									12	25	23	41
	2003									11	23	24	42
	2004									12	23	23	42
	2005									12	23	20	45
	2006	8	22	28	41	17	17	28	39	12	21	30	37
	2007	9	22	28	41	16	16	28	40	12	21	30	37
	2008	7	16	33	44	11	19	27	42	10	27	23	40
	2009	6	15	33	45	6	16	32	46	10	26	23	42
	2010	6	15	33	46	5	15	32	48	8	25	23	44
	2011	5	13	30	52	4	15	27	54	7	23	28	41
	2012	6	15	30	50	5	13	29	53	10	23	24	43
	2013	6	15	30	48	5	14	30	51	11	23	24	42
8	1999	15	22	30	33	15	25	25	35				
	2000	13	24	33	30	18	20	21	41				
	2001	17	26	33	24	17	19	18	45				

Grade	Year	Reading				Mathematics				Science			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	2002	17	23	34	25	16	19	20	46				
	2003	19	27	31	24	16	17	18	48				
	2004	16	24	35	25	14	18	18	50				
	2005	12	25	35	28	15	18	19	48				
	2006	8	25	34	33	12	18	24	47				
	2007	8	26	33	32	11	17	24	48				
	2008	10	19	30	41	12	15	26	47				
	2009	10	19	30	42	11	14	26	49				
	2010	8	18	30	44	10	13	26	51				
	2011	6	16	36	42	9	12	27	52				
	2012	6	14	34	46	6	13	27	54				
	2013	6	14	34	45	6	14	27	53				

Note 1: Percentages in each row may not total exactly to 100% due to rounding.
Note 2: The norm of 2006 and forward is based on the SAT 10 norms and 1999 through 2005 norms are based on SAT 9 norms.
Note 3: Starting in 2008, reading includes the ELL group.

Correlations between Subjects

Table 5.4 shows the correlations among content subjects at each grade level. The correlations are computed using scale scores and Pearson r coefficients. The correlations range from .74 to .81 across grades and content areas. Table 5.5 shows sample sizes involved in the computation of the correlations.

Table 5.4: Correlations among ISAT Scale Scores

Grade	Subject	Subject/Correlation		
		Reading	Mathematics	Science
3	Reading	1.00	0.75	
	Mathematics	0.75	1.00	
4	Reading	1.00	0.77	0.81
	Mathematics	0.77	1.00	0.78
	Science	0.81	0.77	1.00
5	Reading	1.00	0.74	
	Mathematics	0.74	1.00	
6	Reading	1.00	0.77	
	Mathematics	0.77	1.00	
7	Reading	1.00	0.77	0.81
	Mathematics	0.77	1.00	0.80
	Science	0.81	0.80	1.00
8	Reading	1.00	0.74	
	Mathematics	0.74	1.00	

Table 5.5: Sample Size of Correlation Computation

Grade	N		
	Reading-Mathematics	Reading-Science	Mathematics-Science
3	151653		
4	142743	142348	142682
5	143240		
6	147574		
7	148867	148362	148621
8	146131		

REFERENCES

Abedi, J. (1997). *Dimensionality of NAEP subscale scores in mathematics* (CSE Technical Report 428). http://www.cse.ucla.edu/CRESST/pages/reports.htm.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Arce-Ferrer, A. (2008). Comparing screening approaches to investigate stability of common items in Rasch test equating. *Journal of Applied Measurement*. *9*(1), 57-67

Arce-Ferrer, A., & O'Neil, T. (2012). *Investigating anchor set purification in test equating*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Case, B. J. (2003). *Universal design* (Pearson Policy Report). San Antonio, TX: NCS Pearson, Inc.

Clark, L. A., & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309-319.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Pearson Brace Jovanovich, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). *Classics in the history of psychology*. http://psychclassics.yorku.ca/cronbach/construct.htm.

Divgi, D. R. (1980). *Dimensionality of binary items: Use of a mixed model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston MA.

Doran, N. J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139-164.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50* (1), pp. 1-73.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices*. Springer-Verlag. New York.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 171-196). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Erlbaum Associates.

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.

Wells, C., Hambleton, R., & Meng, Y. (2011). *An examination of two procedures for identifying consequential item parameter drift*. (Center for Educational Assessment Research Report No. 761.) Amherst, MA: University of Massachusetts, Center for Educational Assessment.

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

Yen, W. M. (1986). The choice of scale for educational measurement: an IRT perspective. *Journal of Educational Measurement, 23*, 299-325.

APPENDIX A: Conditional Standard Errors of Measurement for ISAT Scale Scores

Conditional SEM (SE_{SS}) for ISAT Reading Scale Scores

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
Raw Score	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}
0	120	47	120	46	120	46	120	47	120	46	120	46
1	120	19	120	20	120	25	120	26	120	24	120	26
2	120	12	120	14	120	16	120	20	120	19	120	19
3	120	11	120	14	120	14	120	18	129	17	124	18
4	120	11	120	14	120	14	127	16	138	15	133	16
5	120	11	123	13	123	13	134	15	145	13	140	14
6	120	11	129	12	129	12	140	14	151	12	146	13
7	120	11	134	11	134	11	145	13	156	11	151	13
8	120	11	138	11	138	11	150	13	161	11	156	12
9	124	11	142	10	143	11	156	12	166	10	161	11
10	128	11	148	10	150	10	167	11	172	10	175	10
11	134	10	153	9	159	9	174	10	177	9	185	9
12	140	10	157	8	167	9	180	9	182	9	193	8
13	145	9	161	8	174	8	185	8	187	8	199	8
14	149	8	165	8	179	8	189	8	191	8	204	8
15	153	8	168	8	184	8	193	8	195	8	207	8
16	157	8	171	8	188	8	196	8	198	8	211	8
17	160	8	175	8	191	7	200	8	201	8	214	8
18	164	8	176	8	193	7	202	8	203	8	216	8
19	167	8	179	8	197	7	205	8	207	8	218	8
20	170	8	181	7	200	7	208	8	210	8	221	8
21	173	8	184	7	203	7	210	8	213	8	223	8
22	176	8	186	7	205	7	213	8	215	8	225	8
23	179	8	188	7	208	8	215	8	218	8	227	8
24	182	8	191	7	210	8	217	8	220	8	229	8
25	184	8	193	7	212	8	219	8	222	8	230	8
26	187	8	195	7	214	8	221	8	224	8	232	8
27	189	8	197	7	216	8	223	8	226	8	234	8
28	192	8	199	7	218	8	225	8	228	8	235	8
29	194	8	202	7	220	8	227	8	230	8	237	8
30	196	8	204	8	222	8	228	8	231	8	238	8
31	198	8	206	8	224	8	230	8	233	8	239	8
32	201	8	208	8	226	8	232	8	235	8	241	8
33	203	8	210	8	228	8	234	8	237	8	242	8
34	205	8	212	8	229	9	236	8	239	8	244	8
35	207	8	215	8	231	9	237	8	240	8	245	8
36	209	9	217	8	233	9	239	8	242	8	247	8
37	211	9	219	8	235	9	241	8	244	8	248	8
38	213	9	221	8	237	9	243	8	246	9	250	9
39	215	9	224	8	239	9	245	8	248	9	251	9
40	217	9	226	8	242	9	247	8	250	9	253	9
41	220	10	229	9	244	9	249	9	252	9	255	9
42	222	10	231	9	246	9	251	9	254	9	256	9
43	224	10	234	9	249	10	253	9	256	9	258	9

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
44	226	10	237	10	252	10	255	10	258	10	260	10
45	229	10	240	10	254	10	258	10	261	11	262	10
46	232	10	243	10	258	11	260	10	264	11	264	10
47	236	10	249	10	261	11	263	10	267	11	267	10
48	238	11	251	11	264	11	267	11	271	12	269	11
49	242	11	256	11	268	11	269	11	274	12	271	11
50	246	12	262	13	273	12	273	12	279	12	275	11
51	252	13	268	14	278	14	277	12	284	13	279	11
52	259	14	278	16	285	15	283	13	292	14	283	12
53	271	17	291	20	293	18	290	14	302	16	290	13
54	288	24	311	29	305	21	301	18	318	21	300	16
55	312	37	332	40	327	32	322	26	346	35	324	24
56	329	47	341	46	351	46	360	47	369	47	379	47

Conditional SEM (SESS) for ISAT Mathematics Scale Scores

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
Raw Score	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}
0	120	49	120	48	120	48	120	48	120	49	120	49
1	120	17	120	20	120	22	120	30	120	30	132	31
2	120	11	120	14	120	18	123	22	130	22	150	22
3	120	10	120	13	125	17	136	18	142	18	163	18
4	120	10	120	12	133	14	145	16	151	15	172	16
5	120	10	124	11	140	12	152	13	158	13	179	14
6	120	10	129	11	145	12	157	12	164	12	185	13
7	120	10	134	11	150	12	162	12	169	11	190	12
8	121	10	138	11	154	11	166	11	174	11	194	11
9	125	10	142	10	158	10	170	11	177	11	198	11
10	129	9	145	10	162	9	173	11	181	10	201	10
11	132	9	149	10	165	9	176	10	184	10	205	9
12	135	9	152	9	168	9	179	10	187	10	208	9
13	139	9	156	9	171	8	182	9	191	10	210	9
14	142	8	161	8	177	8	184	8	195	9	213	9
15	145	8	164	7	182	7	187	8	199	8	215	8
16	148	8	168	7	187	7	190	8	202	8	218	8
17	151	8	171	7	191	7	195	8	206	8	220	8
18	154	7	174	7	195	7	199	7	209	8	222	8
19	157	7	177	7	198	7	203	7	213	7	224	8
20	160	7	179	7	201	7	207	7	216	7	226	8
21	163	7	182	7	205	7	211	7	219	7	229	7
22	166	7	184	7	207	7	214	7	221	7	231	7
23	169	7	187	7	210	7	217	7	225	7	232	7
24	171	7	189	7	213	6	220	6	228	7	234	7
25	173	7	191	7	215	6	223	6	231	7	236	7
26	177	7	193	7	217	6	225	6	234	7	238	7
27	179	7	195	6	219	6	228	6	236	7	240	7
28	181	7	197	6	222	6	230	6	239	6	242	7
29	184	7	199	6	224	6	233	6	241	6	244	7
30	186	7	201	6	225	6	235	6	244	6	246	7
31	188	7	203	6	227	7	238	6	246	6	247	7
32	190	7	205	6	229	7	240	6	248	6	249	6
33	192	7	206	7	231	7	242	6	250	6	251	6
34	194	7	208	7	233	7	244	6	252	6	253	6
35	196	7	210	7	235	7	247	7	254	6	255	6
36	198	7	212	7	236	7	248	7	256	7	257	6
37	200	7	214	7	238	7	250	7	257	7	259	6
38	202	7	216	7	240	7	252	7	259	7	261	6
39	204	7	217	7	241	7	254	7	261	7	262	6
40	206	7	219	7	243	7	256	7	263	7	264	6
41	208	7	221	7	244	7	258	7	264	7	267	6
42	210	7	224	7	246	7	260	7	266	7	268	6
43	212	7	225	7	248	7	262	8	268	7	270	6
44	214	7	226	7	249	7	264	8	269	7	272	6
45	216	7	228	7	251	8	265	8	271	7	274	6
46	217	7	230	7	253	8	267	8	272	7	275	7
47	219	8	232	8	254	8	269	8	274	7	277	7
48	221	8	234	8	256	8	271	8	275	7	279	7

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
Raw Score	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}	Scale Score	SE_{SS}
49	223	8	236	8	258	8	273	9	277	7	281	7
50	225	8	237	8	260	8	274	9	279	8	283	7
51	227	8	239	8	261	8	276	9	280	8	284	7
52	229	9	241	8	263	8	278	9	282	8	286	7
53	231	9	243	8	265	8	280	9	284	8	288	7
54	233	9	245	8	267	9	282	9	285	8	290	7
55	236	9	247	8	269	9	284	9	287	8	292	8
56	238	9	249	8	270	9	285	10	289	8	294	8
57	240	9	251	9	272	10	287	10	290	8	296	8
58	242	10	252	9	274	10	289	10	292	8	298	8
59	244	10	254	9	276	10	292	11	294	9	300	8
60	247	10	256	10	279	10	294	11	296	9	303	8
61	249	11	259	11	280	10	296	11	298	9	305	9
62	252	11	261	11	283	11	298	12	300	10	307	9
63	255	11	263	11	285	11	300	12	302	10	310	10
64	257	12	265	11	288	12	303	13	305	10	313	10
65	260	12	267	11	291	13	305	13	307	11	316	11
66	263	13	270	11	293	13	308	14	310	11	319	11
67	266	13	273	12	296	14	311	15	312	11	322	11
68	270	14	276	13	299	15	314	15	315	12	326	11
69	274	15	279	13	303	16	318	17	319	13	330	12
70	278	16	282	13	307	17	322	18	323	14	335	13
71	283	17	287	14	312	19	328	20	327	15	341	15
72	290	21	292	16	319	22	334	23	332	16	347	17
73	298	25	298	19	327	25	342	26	339	18	356	19
74	311	32	309	23	338	30	351	31	349	23	368	24
75	326	41	326	32	349	36	363	38	363	31	385	33
76	341	49	355	48	369	48	379	48	392	48	410	48

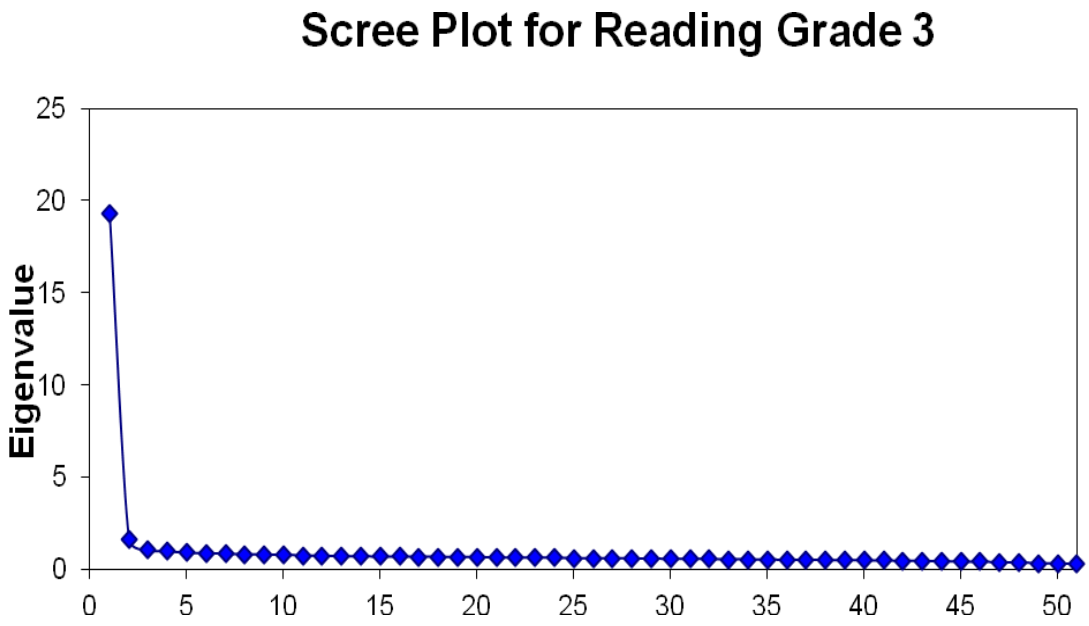
Conditional SEM (SE_{SS}) for ISAT Science Scale Scores

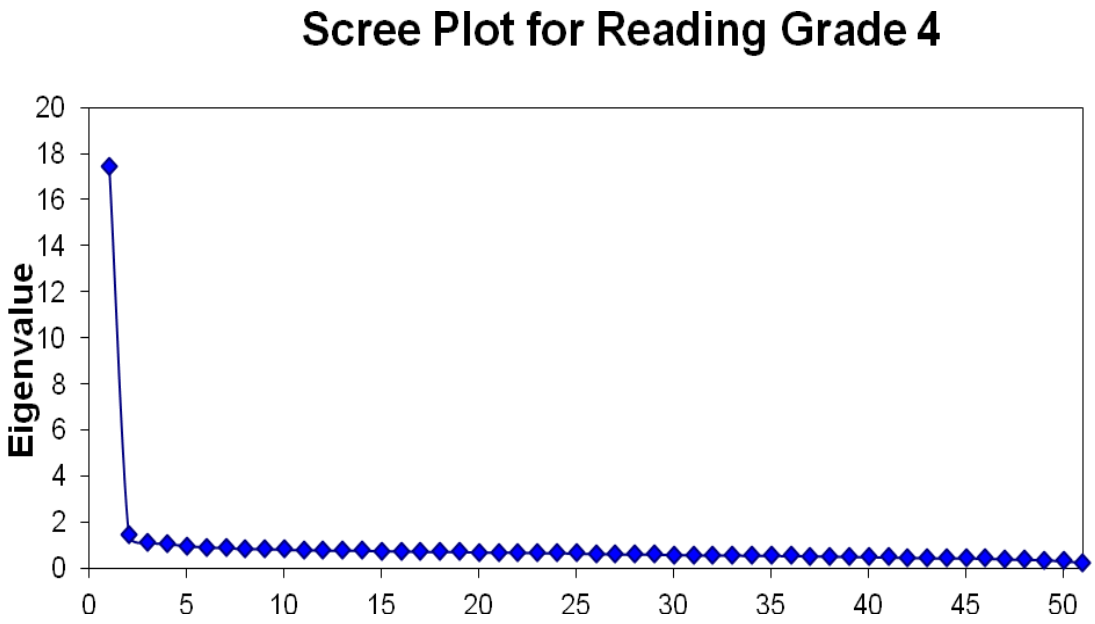
	Grade 4		Grade 7	
Raw Score	Scale Score	SE_{SS}	Scale Score	SE_{SS}
0	120	55	120	55
1	120	31	120	31
2	120	22	120	22
3	120	18	120	18
4	120	16	120	16
5	120	14	125	14
6	120	13	131	13
7	120	12	137	12
8	120	11	142	11
9	120	11	146	11
10	120	11	150	11
11	124	10	153	10
12	127	10	157	10
13	131	10	160	10
14	134	9	163	9
15	136	9	165	9
16	139	9	168	9
17	142	9	171	9
18	144	8	173	8
19	146	8	175	8
20	149	8	178	8
21	151	8	180	8
22	153	8	182	8
23	155	8	184	8
24	158	8	186	8
25	159	8	188	8
26	161	8	191	8
27	164	8	192	8
28	165	8	194	8
29	167	8	197	8
30	169	8	198	8
31	171	8	200	8
32	173	8	202	8
33	175	8	204	8
34	177	8	206	8
35	178	8	207	8
36	180	8	209	7
37	182	8	211	7
38	184	8	214	7
39	186	8	215	7
40	187	8	217	8
41	190	8	218	8
42	191	8	220	8
43	193	8	222	8
44	195	8	224	8

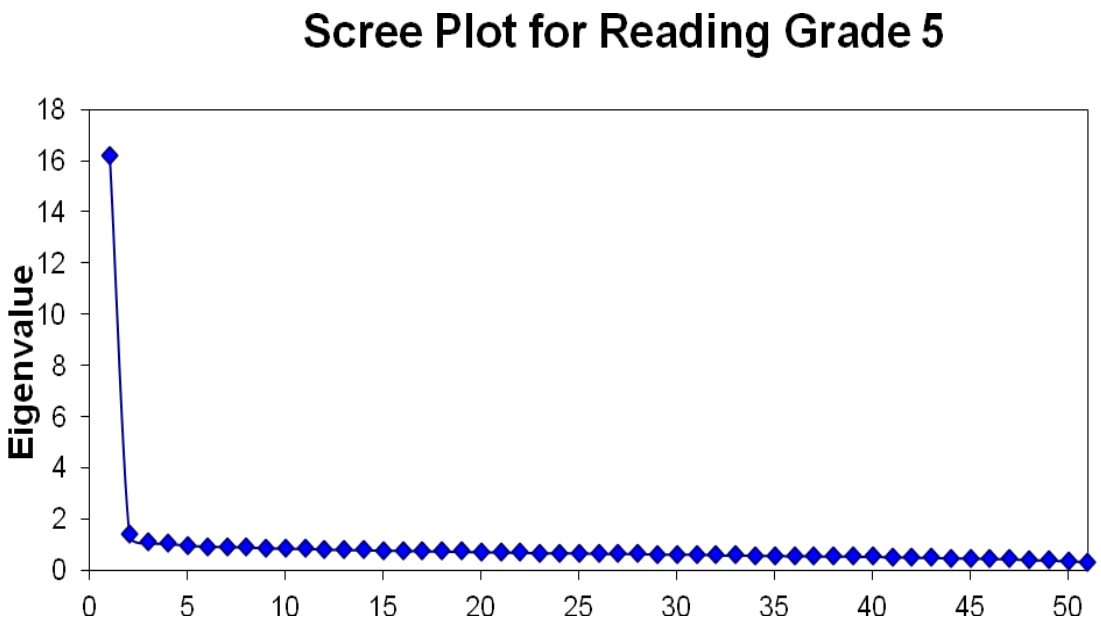
	Grade 4		Grade 7	
Raw Score	Scale Score	SE_{SS}	Scale Score	SE_{SS}
45	197	8	226	8
46	199	8	227	8
47	201	8	230	8
48	203	8	231	8
49	205	8	233	8
50	207	8	235	8
51	209	8	237	8
52	211	8	240	8
53	213	8	242	8
54	215	8	244	8
55	217	8	246	8
56	220	8	248	8
57	222	8	251	8
58	224	9	253	9
59	227	9	256	9
60	230	9	258	9
61	233	9	260	9
62	235	10	264	10
63	237	10	267	10
64	242	10	270	10
65	245	11	274	11
66	249	11	278	11
67	253	11	282	11
68	258	12	287	12
69	263	13	292	13
70	270	14	298	14
71	277	16	306	16
72	286	18	315	18
73	299	22	328	22
74	321	31	350	31
75	361	55	390	55

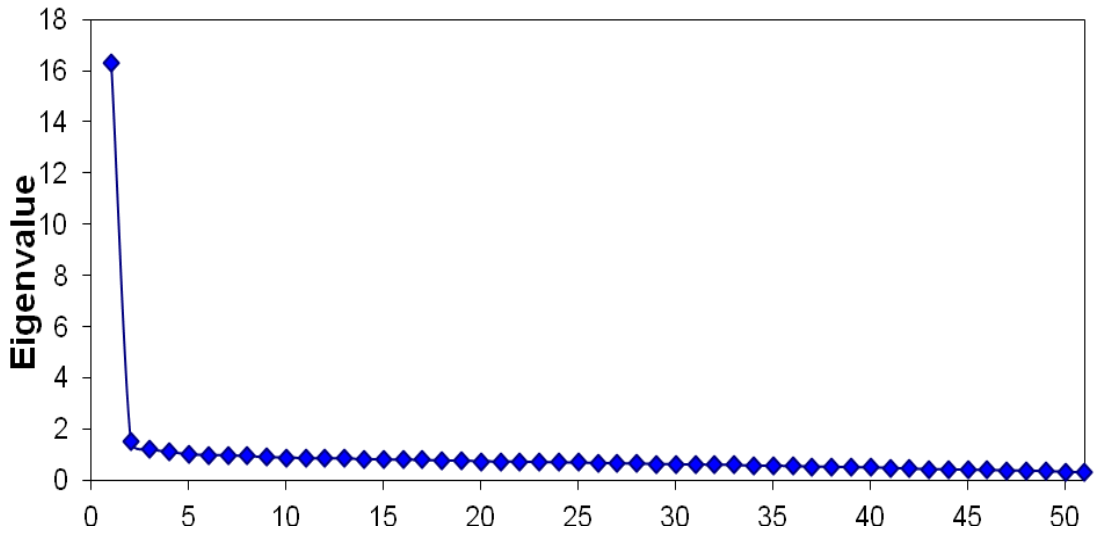
APPENDIX B: Dimensionality Study Scree Plots

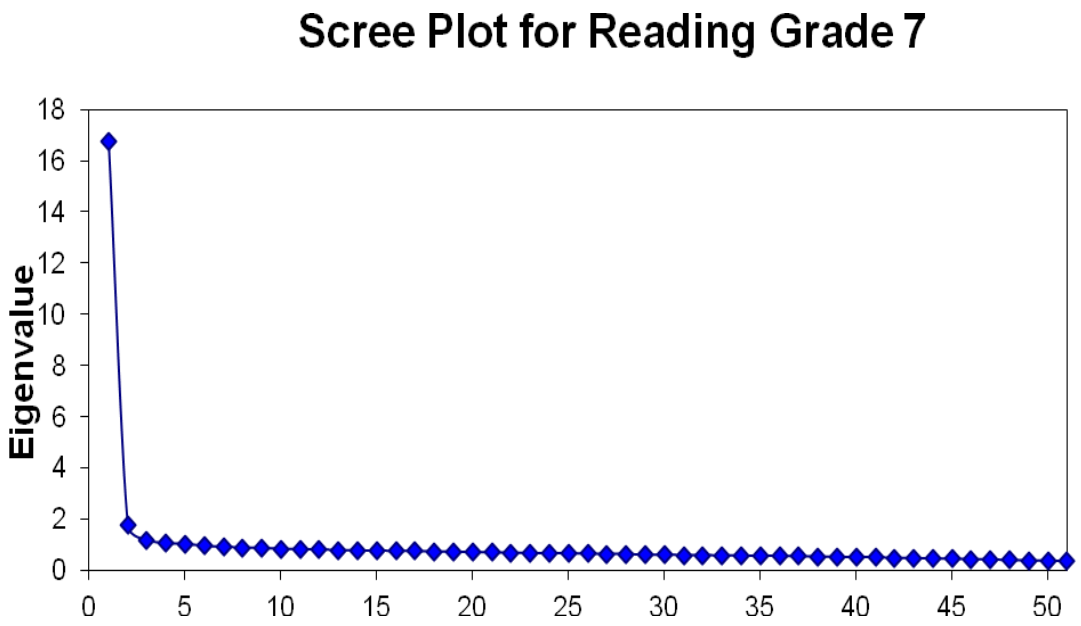
Exploratory Factor Analysis Scree Plots for Reading

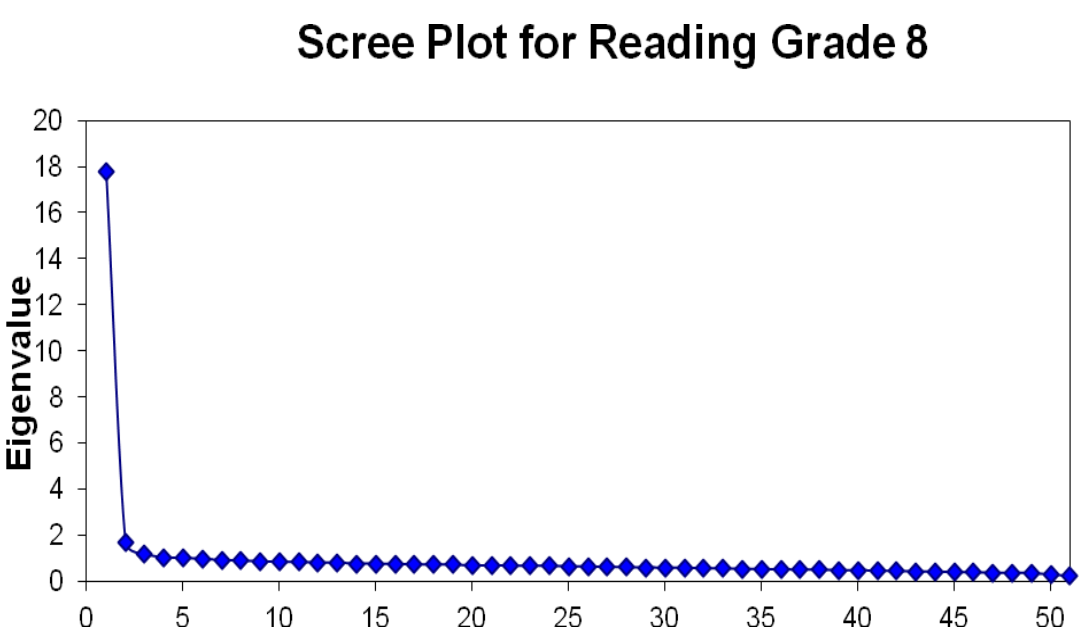




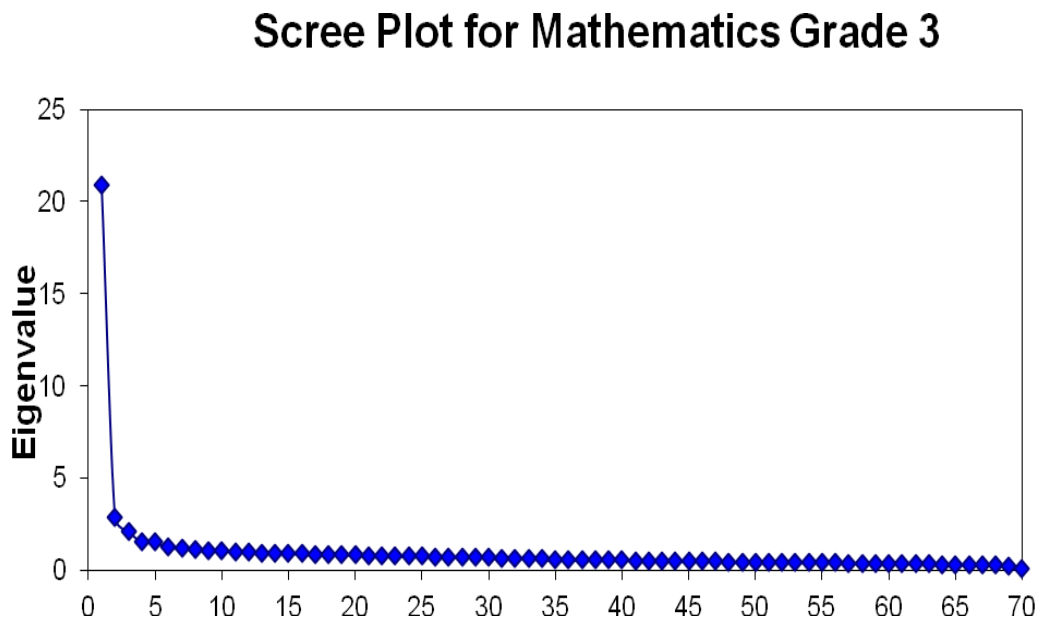

Scree Plot for Reading Grade 5
Eigenvalue
18
16
14
12
10
8
6
4
2
0
0
5
10
15
20
25
30
35
40
45
50

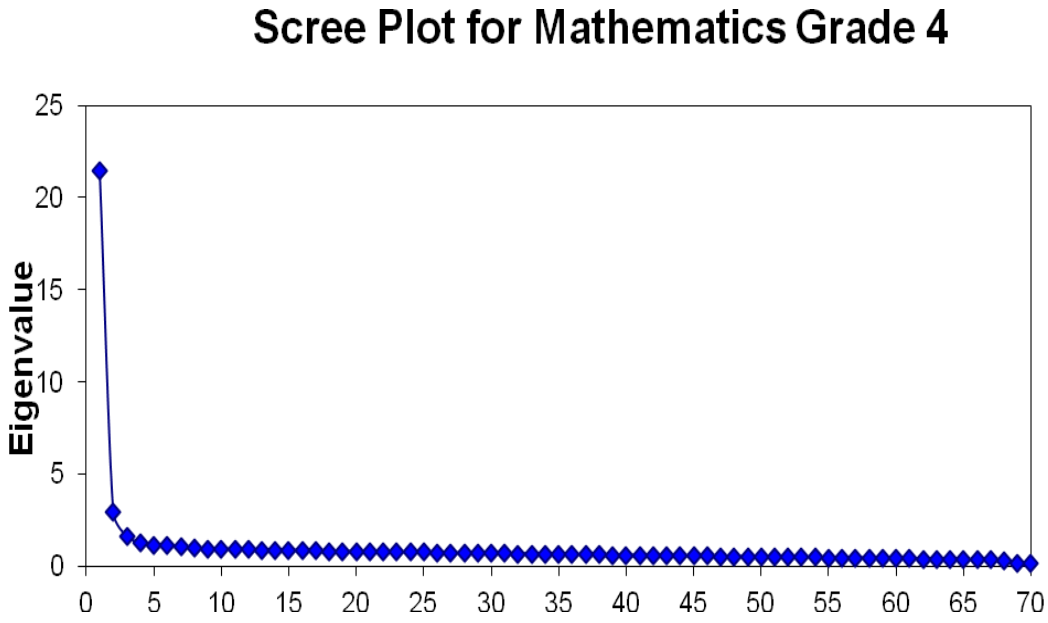

Scree Plot for Reading Grade 6
Eigenvalue
18
16
14
12
10
8
6
4
2
0
0
5
10
15
20
25
30
35
40
45
50


Scree Plot for Reading Grade 7
Eigenvalue
18
16
14
12
10
8
6
4
2
0
0
5
10
15
20
25
30
35
40
45
50

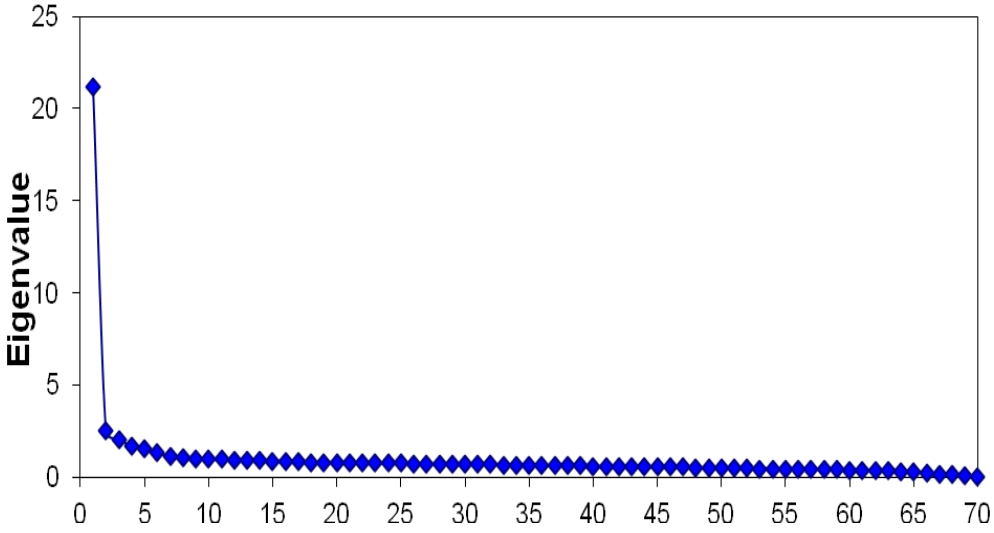

Scree Plot for Reading Grade 8
Eigenvalue
20
18
16
14
12
10
8
6
4
2
0
0
5
10
15
20
25
30
35
40
45
50

Exploratory Factor Analysis Scree Plots for Mathematics

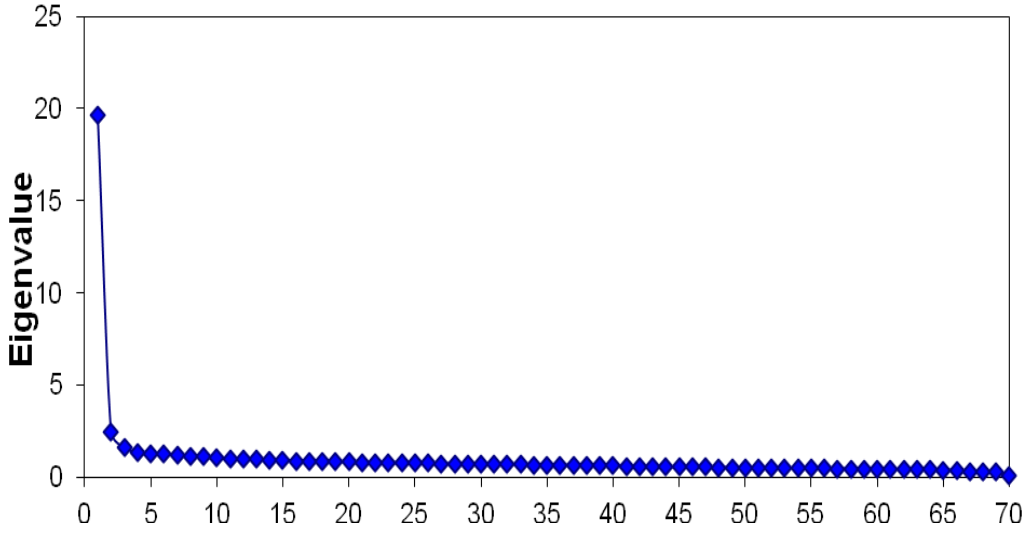




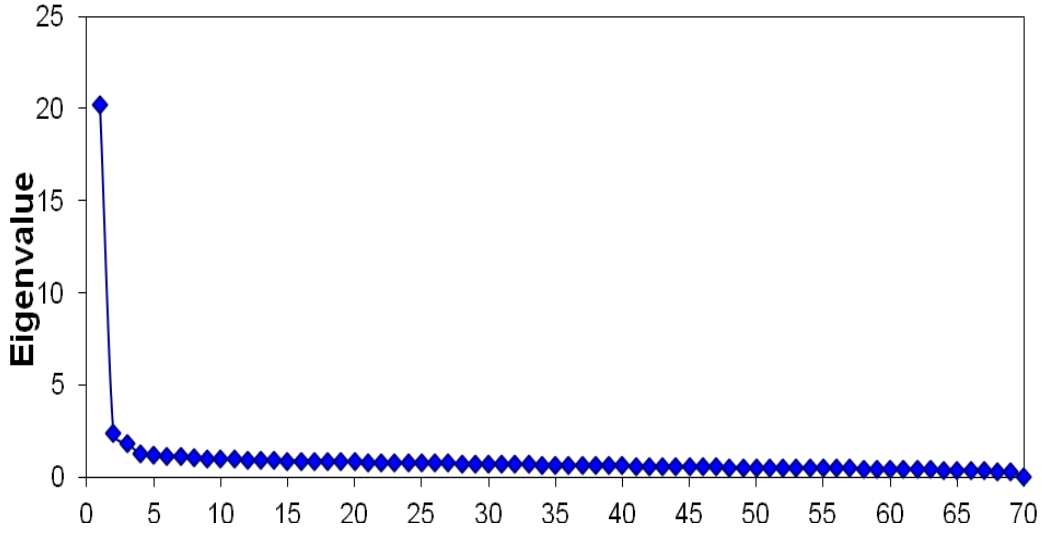
Scree Plot for Mathematics Grade 5



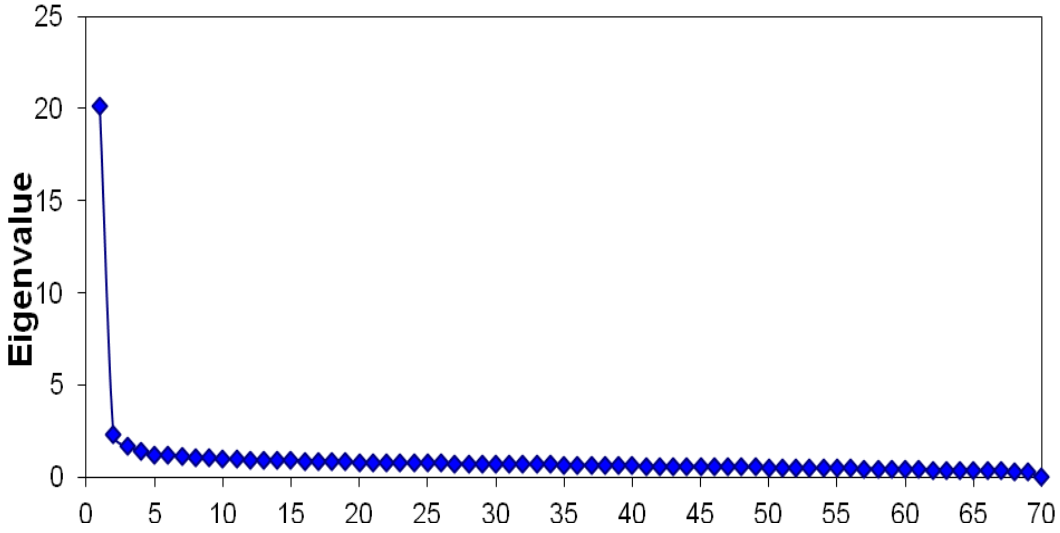
Scree Plot for Mathematics Grade 6



Scree Plot for Mathematics Grade 7



Scree Plot for Mathematics Grade 8



Exploratory Factor Analysis Scree Plots for Science