HiSET®

Measuring the Power of Learning.®

HiSET® Technical Manual

hiset.org

Table of Contents

Introduction

HiSET® Technical Report

List of Tables

List of Figures

Executive Summary

This technical report documents the development, delivery, analyses, and results of the *HiSET®* battery of assessments and presents an analysis of the data from the 2015 HiSET administration. Educational Testing Service (ETS) and Iowa testing Programs (ITP) jointly developed the HiSET battery that consists of:

- Language Arts — Reading,
- Language Arts — Writing,
- Mathematics,
- Science, and
- Social Studies.

The HiSET subtests assess the foundational core of academic skills that represent the long-term goals of secondary education, particularly the critical thinking skills of analysis and evaluation. The HiSET subtests are based on the College and Career Readiness Standards (CCRS) for adult learners (Pimentel, 2013; https://lincs.ed.gov/publications/pdf/CCRStandardsAdultEd.pdf). The CCRS describe the skills and knowledge that adults and youth who have not graduated from high school should acquire to successfully be prepared to enter a job, a training program, or an entry-level, credit-bearing postsecondary course. While the emphasis on particular skills may differ from job to job and course to course, mastery of a core set of essential skills is required.

The results of the HiSET exam are used to determine test taker performance in relation to:

1. The level of academic skills and knowledge typically required to earn a high school equivalency credential, and
2. The level of academic skills necessary to be successfully prepared to enter a job, a training program, or an entry-level, credit-bearing postsecondary course (i.e., college and career ready).

The high school equivalency credential is issued by the state or jurisdiction in which the test taker resides. Depending upon the jurisdiction/state, the high school equivalency credential can be a high school equivalency certificate, high school equivalency diploma, or other documentation as determined by the issuing jurisdiction/state.

This technical report includes the following topics:

- Description of the HiSET program,
- Test design and development,
- Test administration,
- Item scoring,
- Classical item analyses and differential item functioning,
- Reliability,
- Validity of score interpretation,
- Establishment and maintenance of score scales,
- Test taker performance, and
- Quality control procedures.

The design of the HiSET battery of assessments follows the content specifications for each HiSET subtest. These knowledge levels were established based on data collected on high school equivalency standards as well as measures that determine test taker progress toward college and career readiness (i.e., CCRS).

The five HiSET subtests are available for administration on paper, as well as on computer. The tests can be administered in English or in Spanish; accommodated forms are available for test takers with special needs. The Reading subtest consists of 40 multiple-choice (MC) items. The Writing subtest consists of 50 MC items and one essay. The Mathematics, Science, and Social Studies subtests each consist of 50 MC items. Each number-correct score on each subtest converts to a corresponding value on a 1 – 20 reported scale score.

Following the 2015 administration classical item analyses and differential item functioning analyses were performed on the data from each HiSET subtest to evaluate the psychometric characteristics of the test items. The item response theory (IRT) three-parameter logistic model (3PL) was used for item calibrations and scaling.

Performance on the HiSET battery results in three performance level classifications:

- Did not pass high school equivalency,
  - Test taker demonstrates minimal understanding of the subject and has not demonstrated the ability to apply the knowledge and skills that are associated with high school graduation requirements.
- Passed high school equivalency,
  - Passed high school equivalency, but not College and Career Ready — Test taker demonstrates adequate understanding of the subject and has the ability to apply the knowledge and skills that are associated with high school graduation requirements.
- Passed college and career readiness,
  - College and Career Ready — Test taker demonstrates thorough understanding of the subject and has the ability to apply the knowledge and skills that are associated with readiness for college and various career paths.

A scale score of at least 8 on each of the five MC HiSET subtests, a score of at least 2 out of 6 on the essay portion of the Writing test, and a combined score on all five subtests of at least 45 are required to pass the HiSET battery and be certified as performing at a level consistent with high school completion equivalency. A scale score of at least 15 on each of the five MC subtests and a score of at least 4 out of 6 on the essay component of the Writing test are required to demonstrate college and career readiness.

Chapter 1: Introduction

1.1 Description of the *HiSET*® Program

The HiSET program is a high school equivalency testing program for youth and adults who did not graduate from high school. Educational Testing Service (ETS) and Iowa testing Programs (ITP), in partnership with state assessment directors, developed the HiSET program to align with the College and Career Readiness Standards (CCRS) for adult learners (Pimentel, 2013; https://lincs.ed.gov/publications/pdf/CCRStandardsAdultEd.pdf). The CCRS describe the knowledge and skills that will enable the test taker "to meet real-world demands of postsecondary training and employment (Pimentel, 2013, p. 3). Additionally, the HiSET program has been developed to directly measure the academic skills that typically define high school coursework. A thorough review of the CCRS and the HiSET program was conducted by content experts, test developers, and measurement experts to ensure alignment of the HiSET subtests to the CCRS. ETS also worked with subject matter experts and conducted alignment studies (see Chapter 8 for details). The result is a test with the intended objectives:

(a) consistent with the emphasis found in high school curricula,

(b) meets the CCRS for adult education and the Office of Adult Education Standards, and

(c) measures essential components of the CCRS.

The results of the HiSET exam are used to certify a test taker's attainment of academic knowledge and skills equivalent to those of a high school graduate. The results also help identify areas in which candidates are college- and career-ready and areas in which they need improvement. Successful completion of the HiSET program indicates that individuals have demonstrated that they have attained the knowledge and skills equivalent to a high school graduate, and are eligible to pursue postsecondary education and/or various career paths.

HiSET test takers are assessed in five content areas: Reading (Language Arts — Reading), Writing (Language Arts — Writing), Mathematics, Science, and Social Studies. Descriptions of the specifications for each of the five tests are provided in the *Test at a Glance* document, available for download at: https://hiset.org/s/pdf/HiSET_Test_at_a_Glance.pdf. The Reading, Mathematics, Science, and Social Studies tests comprise multiple-choice (MC) items, while the Writing test contains both MC items and one essay. Table 1 presents the number of items and time limits associated with each subtest. The HiSET subtests are available for year-round, continuous testing and are administered on paper and on computer. The subtests are available in English and Spanish, as well as Braille, Reader Script, Large Print, and Cassette or CD.

Each of the five subtests in the HiSET battery is scored on a scale of 1–20. In order to pass, a test taker must do all three

- Achieve a scaled score of at least 8 on each of the five subtests,
- Score at least 2 out of 6 on the essay portion of the Writing test, and
- Have a total combined score on all five subtests of at least 45.

Some states may set passing scores that are higher than this, but under no circumstances can a test taker pass and be certified as performing at a level consistent with high school equivalency with a total score lower than 45 on the full battery of tests. The HiSET tests also results in a College and Career Readiness (CCR) score. A CCR scale score of at least 15 out of 20 on each multiple-choice test and at least 4 out of 6 on the essay are required to demonstrate college and career readiness.

Table 1.1 Number of Items and Time Limits

HiSET Subtest	Number of MC Items	Number of Minutes
Reading	40	65
Writing	50 + 1 essay	75 + 45
Mathematics	50	90
Science	50	80
Social Studies	50	70

1.2 Appropriate Use of Test Scores and Performance Levels

Once the tests are administered, scale scores (total test) and pass/fail decisions are generated for each subtest, and performance is reported at the individual test taker and state levels. The subtest score is used to determine test takers' performance levels, indicating whether or not they passed the subtest. The pass/fail decision is used to inform the test takers whether they have attained the proficiency of high school equivalent skills and knowledge.

The HiSET program provides an Individual Test Report for each test taker. There is an Individual Test Report for each of the HiSET subtests. (A sample report for Language Arts — Writing is provided at: https://hiset.org/s/pdf/Individual-Test-Report-Sample-Report.pdf.) This report indicates, for each HiSET subtest, the test taker's scale score, the minimum scale score required to pass (i.e., high school equivalency), whether the test taker achieved the minimum scale score to achieve high school equivalency, and whether the test taker demonstrated college and career readiness. Finally, the Individual Test Report provides a performance summary for each content category to identify areas of strength and opportunities to improve. Each time a test taker takes one of the HiSET subtests, they will receive an Individual Test Report.

A Comprehensive Score Report is also available; a sample is provided at: https://hiset.org/s/pdf/Comprehensive-Score-Report-Sample-Report.pdf. This report specifies, for each HiSET subtest, whether the test taker met the three HiSET passing criteria. The report also presents a cumulative record of the highest scale score(s) obtained on each subtest, and whether the test taker passed the HiSET battery.

1.3 Overview of the Technical Report

The technical report is organized as follows:

- Chapter 1 — Introduction
  This chapter provides an overview of the HiSET program.
- Chapter 2 — Test Design and Development
  This chapter describes the content framework, test blueprints, and item and form development.
- Chapter 3 — Test Administration
  This chapter provides a description of the test administration procedures, security procedures, and test accommodations.
- Chapter 4 — Item Scoring
  This chapter describes the scoring process for written essays.
- Chapter 5 — Classical Item Analysis
  This chapter describes the data screening criteria, the various classical item statistics, item flagging criteria, and summary results.
- Chapter 6 — Differential Item Functioning
  This chapter describes the analysis procedure, demographic groups included in the analyses, and summary results.
- Chapter 7 — Reliability
  This chapter provides information on reliability and standard error of measurement estimation, results for subgroups of interest, interrater reliability, and classification accuracy and consistency.
- Chapter 8 — Validity
  This chapter describes validity evidence based on test content and internal test structure, as well as results of speededness analyses.
- Chapter 9 — Establishment and Maintenance of Score Scales
  This chapter provides an overview of the method by which the scale score was developed, test equating using item response theory, and calibration and scaling.
- Chapter 10 — Test Taker Performance
  This chapter describers the scale score and performance level results, as well as information to support interpretation of scores.
- Chapter 11 — Quality Control Procedures
  This chapter provides details of procedures implemented to monitor the quality of test materials, system functionality, psychometric analyses, and scoring and reporting.

All technical support and analyses were carried out in accordance with both the *ETS Standards for Quality and Fairness* (2014a) and the *Standards for Educational and Psychological Testing*, issued jointly by the American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014).

Chapter 2: Test Design and Development

2.1 HiSET Content Framework

The *HiSET®* assessment has been carefully designed, developed, and researched to support the two purposes of (a) determining whether a test taker has demonstrated the appropriate level of academic skills and knowledge typically required to earn a high school credential and (b) determining whether a test taker has demonstrated the appropriate level of academic skills to successfully enter a job, a training program, or a postsecondary education program.

The procedures used to develop and revise the test materials are the foundation for the assessment's content validity. Meaningful evidence related to inferences based on high school content and performance standards has guided the design and development of the content of this assessment.

HiSET has been designed and implemented according to established professional standards in order to ensure that the assessment is a measure of what it claims to be, and to support reliable and valid interpretations of test scores. This was achieved by following the guidelines in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

2.1.1 Content Validity

The content of the HiSET exam is developed through an iterative process during which test materials are developed and administered to representative national samples of test takers in order to evaluate the measurement quality and appropriateness of the materials. The HiSET development process begins with the drafting of test specifications that define the knowledge and skills to be measured from the high school curriculum. Reviews of local, state, and national guidelines (College and Career Readiness Standards [CCR] for Adult Education) for high school curriculum and input of school administrators, curriculum specialists, and classroom teachers help to define the test specifications. Educators at the secondary level are consulted on the importance of the knowledge and skills included in the test and the relative importance of these knowledge and skills. New forms of the assessment will be developed to be consistent with shifts in curriculum and instructional practice as reflected in typical high school coursework.

2.1.2 Fairness

Concern for fairness and the elimination of bias from the assessment is a guiding principle throughout design and development. In particular, the HiSET battery was built with careful attention to content-related sources of test bias. Procedures addressed this source of bias, through the following activities:

1. Thorough examination of content and performance standards for the selection of the appropriate content.
2. Engagement of panels of experts in the review of the test specifications, items, and forms.
3. Alignment of items to the defined test specifications.
4. Statistical procedures for identifying items on these tests that function differently across various groups of test takers.
5. Careful selection of a national sample of test takers to respond to the assessment.

2.2 Test Blueprint

2.2.1 Language Arts — Reading

The Language Arts — Reading test provides evidence of a test taker's ability to understand, comprehend, interpret, and analyze a variety of reading material. In the HiSET program, test takers are required to read a broad range of high-quality, challenging literary and informational texts. The texts reflect multiple genres on subject matter that varies in purpose and style. The selections may be memoirs, essays, biographical sketches, editorials, narrations, or poetry. The texts generally range in length from approximately 400 to 600 words. Table 2.1 shows the content categories and corresponding approximate percentages of items for the Language Arts — Reading test.

Table 2.1 Language Arts – Reading: Content Categories and Distribution of Items

Content Category	Approximate Percentage of Items
I. Literary texts	60%
II. Informational texts	40%

Language Arts — Reading Process Categories

The Process Category Descriptors describe in greater detail the skills and knowledge eligible for testing. Test takers answer questions about the provided texts that may involve one or more of the Process Category Descriptors that are numbered under each Reading Process below:

A. Comprehension
   1. Understand restatements of information.
   2. Determine the meaning of words and phrases as they are used in the text.
   3. Analyze the impact of specific word choices on meaning and tone.

B. Inference and interpretation
   1. Make inferences from the text.
   2. Draw conclusions or deduce meanings not explicitly present in the text.
   3. Infer the traits, feelings, and motives of characters or individuals.
   4. Apply information.
   5. Interpret nonliteral language.

C. Analysis
   1. Determine the main idea, topic, or theme of a text.
   2. Identify the author's or speaker's purpose or viewpoint.
   3. Distinguish among opinions, facts, assumptions, observations, and conclusions.
   4. Recognize aspects of an author's style, structure, mood, or tone.
   5. Recognize literary or argumentative techniques.

D. Synthesis and generalization
   1. Draw conclusions and make generalizations.
   2. Make predictions.
   3. Compare and contrast.
   4. Synthesize information across multiple sources.

2.2.2 Language Arts — Writing

The Language Arts — Writing test provides information about a test taker's skill in recognizing and producing effective standard American written English, or Spanish for Spanish speaking test takers. The multiple-choice items measure a test taker's ability to edit and revise written text. The essay question measures a test taker's ability to generate and organize ideas in writing.

The multiple-choice items require test takers to make revision choices concerning organization, diction, and clarity, sentence structure, usage, and mechanics. The test items are embedded in complete texts which span various forms (e.g., letters, essays, newspaper articles, personal accounts, and reports).

The texts are presented as drafts in which parts have been underlined or highlighted to indicate a possible need for revision. Test items present alternatives that may correct or improve the indicated portions. Table 2.2 shows the content and corresponding approximate percentages of items for the Language Arts — Writing test.

Table 2.2 Language Arts — Writing: Content Categories and Distribution of Items

Content Category	Approximate Percentage of Items
Multiple-Choice Items	
I. Organization of ideas	26%
II. Language facility	44%
III. Writing conventions	30%
Content Category	**Number of Items**
Essay Question	
A. Development of ideas	One essay question
B. Organization of ideas	
C. Language facility	
D. Writing conventions	

Language Arts — Writing Content Categories — Multiple-choice Items

The Content Category Descriptors describe in greater detail the skills and knowledge eligible for testing. Because the Language Arts — Writing assessment was designed to measure the ability to analyze and evaluate writing, answering any item may involve aspects of more than one category. The Content Category Descriptors are numbered under each Content Category below for multiple-choice items, followed by the Content Category Descriptors for the essay question.

I. Organization of ideas
   1. Select logical or effective opening, transitional, and closing sentences.
   2. Evaluate relevance of content.
   3. Analyze and evaluate paragraph structure.
   4. Recognize logical transitions and related words and phrases.

II. Language facility
   1. Recognize appropriate subordination and coordination, parallelism, and modifier placement.
   2. Recognize effective sentence combining.
   3. Recognize idiomatic usage.
   4. Maintain consistency and appropriateness in style and tone.
   5. Analyze nuances in the meaning of words with similar denotations.

III. Writing conventions
   1. Recognize verb, pronoun, and modifier forms.
   2. Maintain grammatical agreement.
   3. Recognize and correct incomplete sentence fragments and run-ons.
   4. Recognize correct capitalization, punctuation, and spelling.
   5. Use reference sources appropriately.

Language Arts — Writing Content Categories — Essay Question

The essay question measures proficiency in the generation and organization of ideas through a direct assessment of evidence-based writing. Test takers read a pair of text passages that are related based on a topic, each presenting a different point-of-view regarding the issue/topic being discussed, and then create written responses. Using the essay scoring rubric, the essay responses are evaluated on the test takers' abilities to develop positions or claims supported by evidence from the materials provided as well as from their own experiences.

The following are descriptions of the skills and knowledge covered in the content categories for the essay question.

A. Development of a Central Position or Claim
   1. Focus on central idea, supporting ideas.
   2. Explanation of supporting ideas.
   3. Command over writing an argument.

B. Organization of Ideas
   1. Introduction and conclusion.
   2. Sequencing of ideas.
   3. Paragraphing.
   4. Transitions.

C. Language Facility
   4. Word choice.
   5. Sentence structure.
   6. Expression and voice.

D. Writing Conventions
   1. Grammar.
   2. Usage.
   3. Mechanics.

2.2.3 Mathematics

The Mathematics test assesses mathematical knowledge and competencies. The test measures a test taker's ability to solve quantitative problems using fundamental concepts and reasoning skills. The test items present practical problems that require numerical operations, measurement, estimation, data interpretation, and logical thinking. Problems are based on realistic situations and may test abstract concepts such as algebraic patterns, precision in measurement, and probability. Table 2.3 shows the content categories and approximate percentages of items for the Mathematics test.

Table 2.3 Mathematics: Content Categories and Distribution of Items

Content Category	Approximate Percentage of Items
I. Numbers and operations on numbers	19%
II. Measurement and geometry	18%
III. Data analysis, probability, and statistics	18%
IV. Algebraic concepts	45%

In addition to knowing and understanding the mathematics content explicitly described in the Content Category Descriptors, test takers also will answer items that may involve one or more of the Process Categories. Each Process Category is further divided into Process Category Descriptors. The Content Category Descriptors are numbered under each Content Category listed below. The Process Category Descriptors are numbered under the Mathematics Process Categories section.

Mathematics Content Description

I. Numbers and Operations on Numbers

1. Know that there are numbers that are not rational, and approximate them by rational numbers. (e.g., identify rational and irrational numbers, locate these numbers between two points on a number line, find the product and sum of rational and irrational numbers, and determine if the product or sum is rational or irrational).
2. Rewrite expressions involving radicals and rational exponents using the properties of exponents.
3. Solve problems using scientific notation.
4. Reason quantitatively and use units to solve problems.
5. Choose a level of accuracy appropriate to limitations on measurement.
6. Solve multistep real-world and mathematical problems involving rational numbers in any form and proportional relationships (settings may include money, rate, percentage, average, estimation/rounding).

II. Measurement/Geometry

1. Use congruence and similarity criteria for triangles to solve problems and to prove relationships in geometric figures.
2. Know properties of polygons and circles, including angle measure, central angles, inscribed angles, perimeter, arc length and area of a sector, circumference, and area.
3. Understand and apply the Pythagorean Theorem.
4. Understand transformations in the plane, including reflections, translations, rotations, and dilations.
5. Use volume formulas for cylinders, pyramids, cones, and spheres to solve problems.
6. Apply concepts of density based on area and volume in modeling situations (e.g., persons per square mile, BTUs per cubic foot).

III. Data Analysis/Probability/Statistics

1. Summarize and interpret data presented verbally, tabularly, and graphically; make predictions and solve problems based on the data. Recognize possible associations and trends in the data.
2. Identify line of best fit.
3. Find the probabilities of single and compound events.
4. Approximate the probability of a chance event, and develop a probability model and use it to find probabilities of events.

5. Use measures of center (mean) to draw inferences about populations including summarizing numerical data sets and calculation of measures of center.
6. Understand how to use statistics to gain information about a population, generalizing information about a population from a sample of the population.

IV. Algebraic Concepts

1. Interpret parts of an expression, such as terms, factors, and coefficients in terms of its context.
2. Perform arithmetic operations on polynomials and rational expressions.
3. Write expressions in equivalent forms to solve problems. Factor a quadratic expression to reveal the zeros of the function it defines.
4. Solve linear equations and inequalities in one variable, including equations with coefficients represented by letters.
5. Solve quadratic equations in one variable.
6. Solve simple rational and radical equations in one variable.
7. Solve systems of equations.
8. Represent and solve equations and inequalities graphically.
9. Create equations and inequalities to represent relationships and use them to solve problems.
10. Rearrange formulas/equations to highlight a quantity of interest.
11. Understand the concept of a function and use function notation; interpret key features of graphs and tables in terms of quantities. Evaluate functions for inputs in their domains, and interpret statements that use function notation in terms of a context. Write a function that describes a relationship between two quantities.
12. Understand domain and range of a function.
13. Write a function that describes a relationship between two quantities, including arithmetic and geometric sequences both recursively and with an explicit formula; use them to model situations, and translate between the two forms.
14. Explain each step in solving a simple equation as following from the equality of numbers asserted at the previous step, starting from the assumption that the original equation has a solution. Construct a viable argument to justify a solution method.
15. Calculate and interpret the average rate of change of a function (presented symbolically or as a table) over a specified interval. Estimate rate of change from a graph.

Mathematics Process Categories

In addition to knowing and understanding the mathematics content explicitly described in the Mathematics Content Description section above, test takers also answer test items that may involve one or more of the processes described below. Any of the following processes may be applied to any of the content areas of the Mathematics subtest:

A. Understand mathematical concepts and procedures
    1. Select appropriate procedures.
    2. Identify examples and counterexamples of concepts.

B. Analyze and interpret information
    1. Make inferences or predictions based on data or information.
    2. Interpret data from a variety of sources.

C. Synthesize data and solve problems
    1. Reason quantitatively.
    2. Evaluate the reasonableness of solutions.

2.2.4 Science

The Science test provides evidence of a test taker's ability to use science content knowledge, apply principles of scientific inquiry, and interpret and evaluate scientific information. Most of the items in the test are associated with stimulus materials that provide descriptions of scientific investigations and their results. Scientific information is based on reports that might be found in scientific journals. Graphs, tables, and charts are used to present information and results.

The science situations use material from a variety of content areas such as physics, chemistry, botany, zoology, health, and astronomy. The test takers may be asked to identify the research question of interest, select the best design for a specific research question, and recognize conclusions that can be drawn from results. Test takers also may be asked to evaluate the adequacy of procedures and distinguish among hypotheses, assumptions, and observations. Table 2.4 shows the content categories and approximate percentages of items for the Science test.

Table 2.4 Science: Content Categories and Distribution of Items

Content Category	Approximate Percentage of Items
I. Life science	49%
II. Physical science	28%
III. Earth science	23%

Science Content Description

The following are descriptions of the topics covered in the basic content categories. Because the assessments were designed to measure the ability to analyze and evaluate scientific information, answering any test item may involve content from more than one process category.

Life science topics may include fundamental biological concepts, including organisms, their environments, and their life cycles; the interdependence of organisms; and the relationships between structure and function in living systems.

Physical science topics may include observable properties such as size, weight, shape, color, and temperature; concepts relating to the position and motion of objects; and the principles of light, heat, electricity, and magnetism.

Earth science topics may include properties of earth materials, geologic structures and time, and Earth's movements in the solar system.

Science Process Categories

In addition to knowing and understanding the science content explicitly described in the Science Content Description section above, test takers also will answer test items that may involve one or more of the processes described below. Any of the following processes may be applied to any of the content topics:

A. Interpret and apply
   1. Interpret observed data or information.
   2. Apply scientific principles.

B. Analyze
   1. Discern an appropriate research question suggested by the information presented.
   2. Identify reasons for a procedure and analyze limitations.
   3. Select the best procedure.

C. Evaluate and generalize
   1. Distinguish among hypotheses, assumptions, data, and conclusions.
   2. Judge the basis of information for a given conclusion.
   3. Determine relevance for answering a question.
   4. Judge the reliability of sources.

2.2.5 Social Studies

The Social Studies test provides evidence of a test taker's ability to analyze and evaluate various kinds of social studies information. The test uses materials from a variety of content areas, including history, political science, psychology, sociology, anthropology, geography, and economics. Primary documents, posters, cartoons, timelines, maps, graphs, tables, charts, and reading passages may be used to present information. The test takers may be asked to distinguish statements of fact from opinion; recognize the limitations of procedures and methods; and make judgments about the reliability of sources, the validity of inferences and conclusions, and the adequacy of information for drawing conclusions. Table 2.5 shows the content categories and approximate percentages of items for the Social Studies test.

Table 2.5 Social Studies: Content Categories and Distribution of Items

Content Category	Approximate Percentage of Items
I. History	35%
II. Civics/Government	35%
III. Economics	20%
IV. Geography	10%

Social Studies Content Description

The following are descriptions of the topics covered in the basic content categories. Because the assessments were designed to measure the ability to analyze and evaluate various kinds of social studies information, answering any test item may involve content from more than one process category.

History content includes historical sources and perspectives; the interconnections among the past, present, and future; and specific eras in U.S. and world history, including the people who have shaped them and the political, economic, and cultural characteristics of those eras.

Civics/Government content includes the civic ideals and practices of citizenship in a democratic society; the role of the informed citizen and the meaning of citizenship; the concepts of power and authority; the purposes and characteristics of various governance systems, with particular emphasis on the U.S. government; and the relationship between individual rights and responsibilities and the concepts of a just society.

Economics content includes the principles of supply and demand, the difference between needs and wants, the impact of technology on economics, the interdependent nature of economies, and how the economy can be affected by governments and how that effect varies over time.

Geography content includes concepts and terminology of physical and human geography; geographic concepts to analyze spatial phenomena and discuss economic, political, and social factors; and interpretation of maps and other visual and technological tools and the analysis of case studies.

Social Studies Process Categories

In addition to knowing and understanding the social studies content described in the Social Studies Content Description section above, test takers also will answer items that may involve one or more of the processes described below. Any of the following processes may be applied to any of the content topics:

A. Interpret and apply
   1. Make inferences or predictions based on data or other information.
   2. Infer unstated relationships.
   3. Extend conclusions to related phenomena.

B. Analyze

1. Distinguish among facts, opinions, and values.
2. Recognize the author's purpose, assumptions, and arguments.

C. Evaluate and generalize

1. Determine the adequacy of information for reaching conclusions.
2. Judge the validity of conclusions.
3. Compare and contrast the reliability of sources.

2.3 Item and Form Development

New forms of the HiSET exam are the result of an extended, iterative process during which test materials are developed and administered to national and state samples to evaluate their measurement quality and appropriateness.

2.3.1 Test Specifications

Test specifications outline (among other attributes) the statistical specifications; distribution of content, skills, and cognitive levels across the test form; test organization; and special accommodations. By establishing these parameters beforehand, the test specifications also help to develop new forms that are as comparable to existing forms as possible. The test specifications provide the blueprint for test construction, defining the necessary steps and procedures. As test development proceeds, the test specifications are continually revisited and evaluated in an iterative process so that the materials available for assembly of the final forms reflect the evolving purposes of the assessments. The test development steps for the HiSET exam are presented in Figure 2.1.

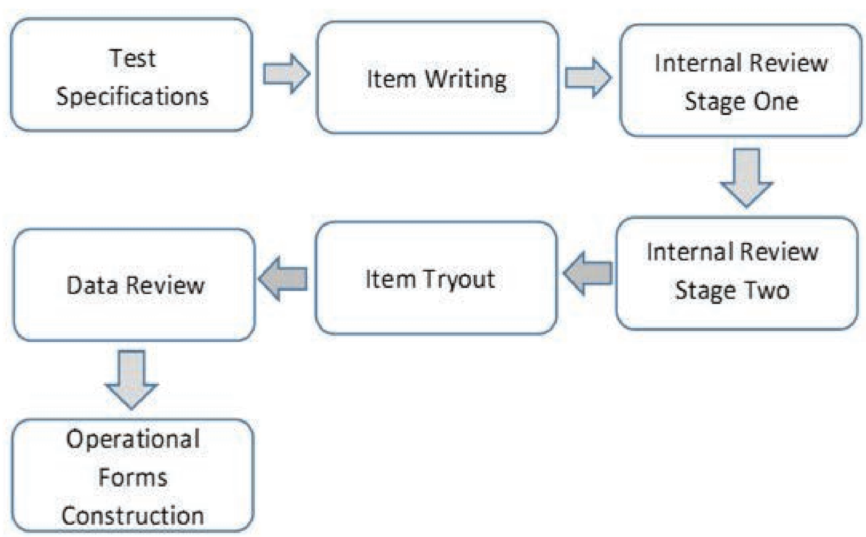

Figure 2.1 Steps in development of the HiSET exam.

2.3.2 Item Writing

Items/item sets and stimuli (reading passages, graphs, maps, tables, and so on that support a group of items) are then created according to the test specifications. HiSET content specialists convene item writing workshops and train educators on sound item writing practices. Educators are assigned to write items in the content areas and grade levels that best align with their experience in the classroom. Item production goals ensure an "overage" of items across content areas for each HiSET subtest so that the pool of available items is far greater than is needed to build the subtests. This overage allows content experts to discard those items that do not survive internal and external item review or post-tryout data review.

After items are written, content specialists review these items for content accuracy, fairness, and universal design (see http://www.cehd.umn.edu/edpsych/C-BAS-R/Docs/Johnstone2008.pdf for an explanation of universal design). The goal of these reviews is to ensure, to the extent possible, that the items are accurate, fair, and accessible to all subgroups in the diverse population of test takers. The items and associated materials are edited to ensure that they are clearly written and that reading loads are appropriate. The items are also copy-edited for grammar and spelling at this stage in the process.

Once the items have been reviewed internally, HiSET content specialists convene panels of educators to review the items and associated stimuli. After a formal training session in the review process, educators review the items for content relevance, and accuracy. Because they have not been involved in the development process up to this point, external reviewers provide an objective "cold read" of potential test materials. A main goal of the educator review is to confirm that the items are appropriate for the intended test takers and HiSET subtest content matter.

HiSET content specialists review the items again after the educator panel review. This review focuses on edits made to the items during previous steps in the process and again checks for content accuracy, fairness, and universal design considerations.

Once items have passed through the review process, data are collected on the performance of the items by conducting a field test to determine how well the items are likely to perform operationally. Test takers complete the field test items when they take the operational tests. It is important that a sufficient number of test takers respond to the field test items to ensure that the associated item statistics are reliable and would accurately reflect the statistics that might be obtained during an operational administration.

The data collected during the field test are analyzed for technical qualities related to item difficulty and item discrimination. This analysis determines whether the items are appropriate measures of test takers' knowledge and the extent to which they will contribute to the test's overall reliability. Only items that display acceptable descriptive statistics are eligible to appear on operational forms. Chapter 5 of this report provides guidelines for acceptable statistical values.

Chapter 3: Test Administration

3.1 Testing Schedule and Administration

This section provides a brief overview of the operational tasks (such as training of test administrators), equipment required, timing instructions, and procedures for implementation of test accommodations for the HiSET test takers.

The *HiSET*® subtests are administered at various test locations, which include numerous community colleges and adult learning facilities. The subtests can be taken on most days when the test centers are open.

The HiSET exam is offered in states/jurisdictions that have adopted the HiSET program. Each state/jurisdiction may have its own requirements for testing, so test takers need to check their state/jurisdiction's requirements before they schedule an appointment to take the test. Test takers can find test centers near them on https://hiset.org/test-takers-hiset-testing-centers/ by entering their city, state and/or ZIP code. Results will display by distance from the center of the location they enter.

Test center staff are trained by ETS on all HiSET administration procedures, related test security issues and the importance of safeguarding test materials.

In addition to ETS's training programs, there are also a number of training manuals and guides that outline everything test center staff need to know to administer the HiSET in compliance with state requirements and ETS policies. Test center staff will receive access to these manuals from ETS. In addition, a number of test administration resources are provided to test centers. Manuals, training modules, and recordings of virtual trainings are available at https://hiset.org/test-centers-administration-resources/. These resources include detailed information on topics such as technology readiness, test administration, test security, accommodations, using the test delivery system, and general testing rules.

3.2 Test Security and Confidentiality

A number of actions are taken to ensure the security of the HiSET program and the confidentiality of test taker information in order to maintain the reliability, validity, and fairness of interpretation of the test results. As mentioned in *Standard 7.9* of *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), "the documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session" (p. 128). Everyone who works with the assessments, communicates test results, and/or receives testing information is responsible for test security, including ETS staff, state assessment coordinators, test center staff, test takers, teachers, and cooperative educational service agency staff. The following paragraphs describe how potential test security incidents are prevented prior to testing and how actual security incidents are handled during and after testing.

The HiSET program developed a test security manual to outline test security responsibilities, expectations, and the process for reporting test security incidents. All educators participating in the administration of the HiSET test are required to participate in the test security training and review the Test Security Manual. Test security training is also incorporated in the on-site test administration workshops. Additionally, test security practices are incorporated into the District/School Assessment Coordinator Guide and the Test Administration Manual (https://hiset.org/s/pdf/HiSET_Program_Manual.pdf).

The State Administrator shall:

1. Inspect each test center before it is established and before approving a change of location.
2. Review emergency plans and test form receiving plans annually for each official HiSET test center in the jurisdiction.
3. Close official HiSET test center(s) when a violation of security procedures occurs and whenever circumstances warrant such action.
4. Oversee investigations of security violations appropriately, including on-site visits whenever feasible.
5. Immediately report any violation of procedures to ETS.

The security of test materials is critical. When test center staff complete all the appropriate steps to establish a HiSET test center, from test administration through the return of test materials to ETS, test center staff are fully responsible for confirming the protection of the tests from loss or unauthorized access. Staff are also responsible for preventing a test taker from having either an unfair advantage or disadvantage. The following procedures must be strictly followed:

1. Make certain no test taker has access to the tests before official test administration.
2. Confirm that every test taker does his or her own work.
3. Verify that no one inspects, views, or reads questions at any time except for test takers when they are taking the test.
4. Test center staff may inspect the content of tests when it is necessary to investigate a test taker's report of a specific problem. Test center staff may read individual test questions only if a test taker reports flawed questions.
5. Based on the ID shown by the test takers, verify that all test takers are authorized to test and that the person taking the test is the person authorized to take it.
6. Provide Test Administrators with a space from which to clearly view all test takers in the testing room at all times.
7. Restrict access to administrative workstation functionalities to authorized test center staff only, and preserve the confidentiality of the information displayed.
8. Notify ETS as soon as possible upon discovery of any potential compromise of test data or materials before, during, or after the testing process.
9. Report any and all unusual testing circumstances by completing a Center Problem Report (CPR). ETS will provide each individual Chief Examiner and/or Test Administrator with his/her own personal login credentials. Personal passwords should never be shared. It is extremely important to protect the integrity and confidentiality of all passwords. A security breach may result in a compromise of the HiSET program and of test taker data.
10. Secure all computers being used for HiSET testing. When test center staff are not present, the testing room must be locked. If a test center uses laptop PCs, then the laptops must be locked in a secure location when not in use.
11. Paper-based testing materials must be secured in a locked room.
12. Any security breach must be reported to the ETS Office of Testing Integrity within 24 hours of the occurrence.

3.3 Reporting Irregularities

No security manual can deal with all situations that might arise during testing. From time to time, questions or emergencies may occur that are not adequately addressed in the manual. ETS relies on test administrators/centers, as the person/entity responsible for all aspects of the administration, to handle any emergency or exceptional situations at the test center. ETS will support test center's actions if they are consistent with established ETS policies and procedures.

The information below provides procedures for documenting testing irregularities and responding to situations that could potentially arise during the course of the test administration.

The guidelines in "Handling Specific Irregularities" are provided as a general framework to facilitate handling of non-routine or emergency situations. ETS staff are available during business hours and on all test dates to offer advice and assistance.

It is extremely important to use the Supervisor's Irregularity Report to report information concerning any possible security breaches, misconduct, and other incidents at the test center to ETS. Facts that may seem of little consequence at the time may later assume considerable significance when ETS staff must decide whether further action is justified.

ETS thoroughly reviews all Supervisor's Irregularity Reports and takes appropriate action. In certain cases, because of confidentiality or privacy factors, it may not be possible for ETS to report back to Test Administrators regarding actions taken.

All reports should be complete and explicit and include a detailed description of the following:

- Overview of the incident or irregularity,
- Identification and appointment number of the individuals involved, including the names and telephone numbers of all test center personnel who might provide relevant information about any tests that might be affected,
- The length of time each incident was observed,
- Details regarding what happened,
- When it happened, and
- The action taken.

A report filed by a Test Administrator should be signed by the Chief Examiner and countersigned by the Test Administrator, who should add any additional information that might also be useful to ETS for resolution of the problem. The report should be completed by test center personnel only.

3.4 Test Accommodations

The HiSET program is committed to serving test takers with disabilities and health-related needs by providing services and reasonable accommodations that are appropriate given the purpose of the test. Accommodations are available for test takers with diagnosed disabilities that include, but are not limited to:

- Attention deficit/hyperactivity disorder,
- Psychological or psychiatric disorders,
- Learning and other cognitive disabilities,
- Physical disorders/chronic health disabilities,
- Intellectual disabilities, and
- Hearing and visual impairment.

Table 3.1 outlines some of the most commonly requested and approved accommodations for paper- and computer-delivered tests. Test takers must request these accommodations prior to scheduling their test appointment. This list includes some, but not all, of the accommodations available to test takers.

Table 3.1 Commonly Approved Accommodations for Paper- and Computer-delivered Tests

Testing Accommodation	Paper	Computer
Extended time	✓	✓
Separate room	✓	✓
Audiocassette or other form of recorded audio	✓	
Braille	✓	
Screen reader		✓
Large print	✓	
Screen magnification		✓
Calculator/talking calculator	✓	✓
Scribe or keyboard entry aide	✓	✓
Additional supervised break time	✓	✓
Sign language-interpreted instructions for deaf or hard-of-hearing test takers	✓	✓

Chapter 4: Item Scoring

4.1 Overview

This chapter documents how ETS Assessment Development (AD) and Performance Assessment Scoring Service (PASS) staff participated in certifying the scoring system and how each team followed procedures required by the ETS Office of Quality for operational readiness and *Standard 7.8* of *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

The writing portion of the HiSET exam:

- Consists of an essay (prompt) section
- Consists of a multiple choice (MC) section

The two sections are scored separately; MC responses are scored by machines; essays are rated by specially trained human raters. The two section scores are converted separately to scaled scores, which are summed to produce the Writing score. This chapter is about the rating (or scoring) of the essay (prompt) section.

4.2 Types of Item Response

Items on the HiSET battery are multiple-choice (MC) except for a direct writing task associated with the Language Arts — Writing subtest. Like other selected-response items, MC items can be answered quickly, making it possible to assess a broad range of content in a limited time. MC items are objectively scored, and the scoring process is quick. However, MC items may not reveal a test taker's reasoning process and typically do not assess higher-order thinking skills.

The writing task on the Language Arts — Writing subtest belongs to the constructed response (CR) category; test takers read a pair of texts which present opposing views on a topic (e.g., should the minimum wage for waiters be increased) and then create a written response on the topic presented. Responses are evaluated on how well the candidates developed positions or claims supported by evidence from the materials provided as well as their own experiences. Test takers can type their essays into a computer or write their essays by hand, depending on which testing mode they choose. Each Writing subtest contains one essay and there are three parallel forms given for the Writing subtest each year, each parallel form has different CR essay question. Essays allow test takers to demonstrate their use of complex thinking skills such as formulating comparisons or contrasts; proposing cause and effects; identifying patterns or conflicting points of view; categorizing, summarizing, or interpreting information; and developing generalizations, explanations, justifications, or evidence-based conclusions. Essays are rated using rubrics written specifically for each prompt, and the essay responses are rated on a 0 to 6 point scale by two raters. The final essay score is the average score assigned by the two raters. In cases where a third rater reviewed the response (explained on the following page), the score from the third rater is used as the final essay score. The essay scores are combined with the MC portion of the Language Arts — Writing subtest to assess language skills.

4.3 Online Scoring and Rater Management

HiSET essays are rated in the ETS Online Network for Evaluation (ONE) system, a distributed, web-based scoring system that enables a large number of raters to view and rate assigned responses from remote locations. All identifying information from the responses sent to the raters is removed so that neither the identity of the test taker nor the test taker's testing center are revealed to the rater; the rater sees only the test taker response. Each essay is rated by two raters to ensure quality of scoring. The second rater rates the response independent of the first rater (e.g., the second rater does not see the score assigned by the first rater). On the 6-point scale, essays scores are considered discrepant if:

a. The scores from the two raters differ by more than **one** point or
b. One rater assigns the response a score of 1 and the other rater assigns the same response a score of 2.

When the essay scores are discrepant, the response is sent to third rater to determine the final score. Note that, responses cannot be given scores of 1 and 2 by the two raters. A score of 1 indicates that the response is not at the high school equivalency level while a score of 2 indicates that the response is high school equivalent. Therefore, the response is sent to third rater for a final rating which is used as the final essay score.

4.3.1 Rater Recruitment and Qualifications

ETS has established procedures for recruiting, training, and certifying raters for online scoring. ETS recruits raters through social media such as LinkedIn and CareerBuilder and through nationwide teachers' associations. Each rater must meet the following minimum requirements:

- have an undergraduate degree from an accredited college or university in the United States,
- reside in the United States,
- be available to work in the United States, and
- be a practicing or former teacher.

Accurate scoring of large numbers of test taker responses requires a comprehensive scoring and leadership structure. The organizational structure for HiSET encompasses four levels of responsibility:

1. Raters. These are the people who rate the responses.
2. Scoring Leaders. Each scoring leader's primary job is to monitor and report on a team of 5 to 10 raters. Scoring leaders "back-read" a sample of responses rated by each of their raters, to see if the raters are applying the scoring rubrics correctly and to correct them if they are not. Scoring leaders are also expected to answer questions raised by raters and to rate non-routine responses. If the scoring leader is unsure of what rating to award a response then the response goes to a group scoring leader or a content scoring leader for resolution.
3. Group Scoring Leaders. These group leaders provide feedback to the scoring leaders while carefully monitoring the overall quality and progress of the scoring by back reading and checking scoring progress in ONE. They rate the complex, non-routine responses and resolve any prompt-related issues raised by the scoring leaders.

4. Content Scoring Leaders. Working under the supervision of ETS AD content experts, content leaders have overall responsibility for one or more of the tests administered by ETS. HiSET content scoring leaders have expertise in the writing content area and work with different Group Scoring Leaders across ETS writing tests. Using ONE scoring report capabilities, they review the performance of the group scoring leaders, and oversee the quality and progress of the scoring. The content scoring leaders work closely with staff from AD, PASS, and Human Resources.

4.3.2 Training

The goal of the rater training is to have all raters apply the same criteria and standards so that the score a response receives will depend as little as possible on which rater rates the response. Raters are trained to apply the scoring rules, as specified in the HiSET scoring rubric and test design. Raters learn to apply the rubrics for each prompt by scoring "benchmark" responses at each rating level. A benchmark response is an actual test taker written response that illustrates the quality expected for responses receiving that rating. The rubric tells the rater what qualities of the essay to consider in rating it, while the benchmark responses indicate, by example, how good an essay has to be to receive each possible score. After completing their training, the raters have to pass a certification test by correctly rating a set of responses that have been previously rated by expert raters. Only after passing the certification test can they begin to rate responses operationally. ETS AD staff conduct the training for raters, using sample responses provided by the HiSET program.

Four types of test taker responses do not receive numeric scores:

- responses that are blank,
- responses written in a language other than the target language (i.e., English or Spanish),
- responses that do not give the rater enough information to assign a valid rating, and
- responses that are "off topic," (i.e., did not reflect an attempt to answer the item).

4.3.3 Certification and Calibration

Before raters are allowed to score student responses they must prove they can apply the rubric correctly, through a process called certification. Certification is the process of determining if a rater has learned the scoring rubric and rating system well enough to apply it. During certification, raters have access to the scoring rubric, benchmark papers exemplifying each score level the rater can assign, and rating notes with information specific to the essay prompt that is being rated. After training on certification materials, raters are provided a set of training papers to practice rating. After raters review all the training papers and practice rating then they complete the certification test. The certification test consists of 10 pre-scored sample responses written to the same prompt and requires the raters to assign scores to the certification responses. ETS staff set the pass threshold for certification. The HiSET minimum passing rate is met by rating 60% of the responses correctly, rating 30% of the responses adjacent to correct, and rating no more than 10% of the responses discrepant. After training, raters are given two chances to pass the certification test. Raters who do not pass on their first certification attempt are given additional practice and a second set of certification responses to rate. Raters who do not pass on their second certification attempt are paid for their practice time but not accepted for rating. This certification structure supports the creation of the pool of qualified raters needed for completion of all rating activities. In addition to the requirements listed above, raters who are inactive for more than 90 days are required to recertify before rating the essay responses.

Similar to the certification sets, calibration sets are a set of responses that have been previously rated by expert raters. Additionally, calibration sets are presented to raters when they are scheduled for a rating session and each calibration set has 10 responses that are rated on a range of 1 to 6 points. Before rating responses to a particular prompt, raters are required to pass at least one of two calibration sets of responses to that prompt. To pass, a rater has to assign the correct rate to at least six of the 10 responses in the calibration set, with no more than three or four scores adjacent and/or one score discrepant. Adjacent scores are scores that are within 1 point of the assigned score; a discrepant score is a score that is 2 or more points away from the assigned score. If the rater is unsuccessful on the first attempt, the rater is required to review the training materials (scoring rubric, benchmark responses, etc.) with the scoring leader and then participate in a second calibration attempt. Raters who do not pass after two calibration attempts are excused from the scoring session.

4.3.4 Quality Control

During rater scoring sessions, ETS creates performance scoring reports so project leadership can monitor the daily scoring process and plan the retraining activities if needed. Scoring reports can indicate which prompts have adjacent or discrepant scores. (Adjacent scores are scores that are within 1 point of the assigned score; a discrepant score is a score that is 2 or more points away from the assigned score.) Scoring leaders are able to monitor scoring performance, while the scoring is going on, with a variety of performance data. To compute performance data, nine percent of the responses assigned to each rater are "monitor responses" or "monitor papers" that have been previously rated by two expert raters. Raters are assigned the monitor papers and data are obtained to show how accurately raters assigned a score to the monitor papers. These monitor responses enable the scoring leader to monitor each rater's accuracy while the rater continues to rate other responses. Although the raters are required to pass the certification tests and the calibration tests prior to scoring, the rater's performance on the monitor papers allows the scoring leader to provide feedback/comments to the rater throughout the rating process.

Chapter 5: Classical Item Analysis

5.1 Overview

This chapter provides a description of the statistical analyses conducted for the HiSET subtests. Classical item analyses involve computing a set of statistics based on the test taker responses for every item in each form. The statistics provide key information about the quality of the items from an empirical perspective. The classical item analyses and the differential item functioning analyses were completed using General Analysis System (*GENASYS*), an ETS proprietary software program. The GENASYS system includes components for establishing test program statistical information (e.g., data layout, number of items, etc.), processing scores for test takers (including case sampling and scoring of multiple-choice items), traditional item analyses, differential item functioning (DIF), item response theory (IRT) analyses, and equating procedures. Using GENASYS, the statistics calculated for the multiple-choice (MC) and constructed response (CR) items, and associated criteria used to identify items that demonstrate less than optimal psychometric characteristics, are described in Section 5.2.

The data sample analyzed for this technical report includes all test takers who took one or more of the HiSET subtests during the 2015 HiSET administration. For each HiSET subtest, test taker records for which there are responses to fewer than five items are excluded from the analyses. Although the HiSET subtests are administered via paper and online in English and Spanish, the statistical analyses described in this Chapter and Chapters 6, 7, 8, and 10 are based on the English online test takers. The English paper forms are printed versions of the English online forms, and the Spanish online and paper forms are direct translations of the English forms. Even though the data from the English paper test takers and the Spanish test takers are not included in the analyses, the Assessment Development group reviews all the forms to ensure the accuracy of the item keys across all the HiSET forms.

5.2 Description of Classical Item Analysis Statistics

1. ***Classical item difficulty indices*** (p-value and average item score). This statistic indicates the mean item score expressed as a proportion of the maximum obtainable item score.

   For MC items, item difficulty is indicated by each item's p-value, which is the proportion of test takers who answered the item correctly. The possible range of p-values for MC items is from 0.00 to 1.00. Items with high p-values are easy items and those with low p-values are difficult items. Desired p-values generally fall within the range of 0.20 to 0.90.

   For CR items, difficulty is indicated by the AIS. The AIS can range from 0.00 to the maximum total possible score for an item (the maximum score for the HiSET essays is six). To facilitate interpretation, the AIS values for CR items are expressed as proportions of the maximum possible score, which are equivalent to the interpretation of p-values for MC items. Desired AIS values generally fall within the range of 20% to 90% of the maximum points possible.

2. ***Item-total correlation of the correct response option***. This statistic measures the strength of the relationship between test takers' performance on a specific item and their performance on the MC portion of each HiSET subtest. For the ELA writing test, the total score does not include the essay. The item-total correlation is bounded by -1.00 and +1.00 and typically ranges from 0.00 to 0.70. Desired values are positive and larger than 0.25. The higher the value, the better the item distinguishes between higher- and lower-scoring test takers. Positive values indicate that the test takers who do well on the test have higher probability of answering the questions correctly, while negative item-total correlations indicate that low ability test takers perform better on an item than high ability test takers. Therefore, negative correlations can indicate serious problems with the item content (e.g., multiple correct answers or unusually difficult or complex content).

   For the MC items, the item-total correlation is the biserial correlation and is computed using the following formula:

$$r_{bis} = \left(\frac{pq}{Y_{zp}}\right)\frac{\left(\bar{x}_1 - \bar{x}_0\right)}{s_{tot}} \tag{5-1}$$

   where p is the proportion of test takers who received a score of 1 on the item,

   q is the proportion of test takers who received a score of 0 on the item,

   Y_{zp} is the Y ordinate (height) of the standard normal curve at the z-score associated with the p-value for the item,

   $\bar{x}_1$ is the total test mean of the test takers who received a score of 1 on the item,

   $\bar{x}_0$ is the total test mean of the test takers who received a score of 0 on the item, and

   s is the standard deviation for the total test.

   For the CR items, the item-total correlation is the polyserial correlation. The polyserial is a generalization of the biserial correlation for items with more than two possible score values (the Writing CRs are scored on a scale of 0 to 6). Polyserial correlations are based on a polyserial regression model (Drasgow, 1988; Lewis & Thayer, 1996), which assumes that performance on an item is determined by the test taker's position on an underlying latent variable that is normally distributed at a given criterion score level. Based on this model, the polyserial correlation can be estimated using the formula:

$$r\ polyreg = \frac{bs}{\sqrt{b^2 s_{tot}^2 + 1}} \tag{5-2}$$

   where b is estimated from the data using maximum likelihood and s_{tot} is the standard deviation of the criterion score.

The polyserial correlation was used because it measures the correlation between two continuous variables, where one variable is observed directly, and the other is unobserved. Information about the unobserved variable is obtained through an observed ordinal variable that is derived from the unobserved variable by classifying its values into a finite set of discrete, ordered values (Olsson, Drasgow, and Dorans, 1982). For HiSET the unobserved variable is derived from the scores on the constructed response items which are scored on a 0 to 6 scale, while the observed continuous variable is the total score on the writing test.

3. ***Percentage of test takers not responding to an item (Speededness)***. This statistic is useful for identifying problems with test features, such as testing time and item/test layout. A not responded to item is classified as either an omitted or a not reached item. If a test taker did not respond to an item, the item is considered to be omitted. An item is considered not reached if the test taker did not respond to that item and any subsequent items. Omit rates for CR items tend to be higher than for MC items. When a pattern of omit percentages exceeds 5 percent for a series of MC items at the end of a timed section, this may indicate that there was insufficient time for test takers to complete all items. For individual items this could be an indication of an item/test layout problem. For example, test takers might accidentally skip an item that follows a lengthy stem.
4. ***Distribution of CR item scores***. For CR items, examination of the distribution of scores is helpful to identify how well the item is functioning. If no test takers' responses are awarded the highest possible rating (a score of six points), this may indicate that the item is not functioning as expected (e.g., the item could be confusing, poorly worded, just unexpectedly difficult, or the test takers may not have understood the writing task). It is possible that the "benchmarks" and/or the "rangefinders" responses/examples that support the scoring of the CRs may be flawed.

5.3 Summary of Classical Item Analysis Flagging Criteria

Flags are letter codes that identify extreme statistical values that may indicate a problem with the item. Flagged items were not removed from subsequent analyses, but the flags served to notify psychometricians and assessment development staff that items were not performing as expected. The following flagging criteria were applied to the MC and CR items:

- *Difficulty flag:* p-values less than 0.20 or greater than 0.90.
- *Discrimination flag:* Item-total correlation less than 0.25.
- *Omit flag:* Percentage of test takers omitting an item greater than 5% for MC items, and greater than 15% for CR items.

5.4 Classical Item Analysis Results

Distributions and summary statistics of the p-values and item-total correlation statistics for all items in the three forms combined, for each subtest, are provided in Table 5.1. Relatively few items were flagged for being very easy or very difficult, with the exception of Mathematics. Mathematics was a difficult subtest with a mean p-value of 0.33 and with 39 items (26%) being flagged as very difficult. In addition, there were more Mathematics items flagged for being difficult or for having a low-item total correlation than for the other four HiSET subtests.

Table 5.1 Summary of *p*-values and Item-total Correlations

	Reading	Writing	Mathematics	Science	Social Studies
Number of Items	120	150	150	150	150
p-value					
≥ 0.90	10	10	1	8	6
0.80 – 0.89	30	24	4	13	12
0.70 – 0.79	22	24	9	33	17
0.60 – 0.69	25	23	6	19	37
0.50 – 0.59	14	33	10	31	36
0.40 – 0.49	13	21	11	23	25
0.30 – 0.39	5	9	19	14	9
0.20 – 0.29	1	6	51	6	7
< 0.20	0	0	39	3	1
Mean	0.69	0.63	0.33	0.60	0.59
Median	0.71	0.64	0.26	0.59	0.59
SD	0.16	0.18	0.20	0.18	0.17
Item-Total Correlation					
≥ 0.60	12	1	8	19	9
0.50 – 0.59	57	27	25	48	37
0.40 – 0.49	35	63	27	42	54
0.30 – 0.39	11	41	35	22	29
0.20 – 0.29	2	11	40	13	14
< 0.20	3	7	15	6	7
Mean	0.50	0.41	0.36	0.45	0.43
Median	0.52	0.42	0.35	0.48	0.45
SD	0.10	0.10	0.14	0.13	0.13

Tables A.1 through A.15 in Appendix A present more detailed results from the classical item analyses for all of the items administered in each form, for each HiSET subtest. These tables provide item statistics and flags, as well as item location information on test forms, for both MC and CR items. These tables also present 3-parameter (3PL) item response theory (IRT) parameter estimates for each MC item. The use of IRT for the HiSET program is described in Chapter 9 of this report. Appendix B, Tables B.1 to B.5 present summaries of the MC item flags, for each form of each subtest (no CR items were flagged). Summaries of *p*-values, item discrimination statistics, and IRT parameter estimates are reported for each test form in Appendix C, Tables C.1 to C.5.

5.5 Speededness

The percentage of test takers who omitted MC and CR items throughout the tests was examined to evaluate whether sufficient time was allowed for the HiSET subtests to be completed. The flagging criteria for high omit rates was more than five percent of test takers omitting an MC item and more than 15 percent of test takers omitting a CR item. Based on these criteria, no MC or CR items were flagged as having high omit rates. As shown in Tables 5.2 through 5.6, almost 100% of the test takers responded to all items across the five subtests.

Table 5.2 Omit and Not Reached Information for Reading

	Form A	Form B	Form C
Number of test takers	9,045	8,910	8,619
Number of items	40	40	40
Percent reaching all items	98.9	98.6	98.7
Percent Reaching 75% of items	99.8	99.8	99.9
Mean number of items omitted (standard deviation)	0.061 (0.631)	0.050 (0.337)	0.058 (0.462)
Mean number of items not reached (standard deviation)	0.068 (1.001)	0.082 (1.064)	0.063 (0.791)

Table 5.3 Omit and Not Reached Information for Writing (MC Items)

	Form A	Form B	Form C
Number of test takers	6,142	9,076	9,005
Number of items	50	50	50
Percent reaching all items	99.6	99.5	99.6
Percent Reaching 75% of items	99.9	99.9	99.8
Mean number of items omitted (standard deviation)	0.070 (0.348)	0.082 (0.424)	0.071 (0.351)
Mean number of items not reached (standard deviation)	0.045 (1.001)	0.045 (0.914)	0.054 (1.247)

Table 5.4 Omit and Not Reached Information for Mathematics

	Form A	Form B	Form C
Number of test takers	10,149	9,937	10,316
Number of items	50	50	50
Percent reaching all items	97.9	98.5	98.2
Percent Reaching 75% of items	99.8	99.8	99.7
Mean number of items omitted (standard deviation)	0.260 (1.414)	0.293 (1.687)	0.307 (1.818)
Mean number of items not reached (standard deviation)	0.121 (1.206)	0.093 (1.230)	0.109 (1.267)

Table 5.5 Omit and Not Reached Information for Science			
	Form A	Form B	Form C
Number of test takers	8,349	8,414	8,235
Number of items	50	50	50
Percent reaching all items	99.5	99.3	98.9
Percent Reaching 75% of items	99.9	99.9	99.9
Mean number of items omitted (standard deviation)	0.087 (0.926)	0.076 (0.771)	0.105 (0.869)
Mean number of items not reached (standard deviation)	0.047 (1.037)	0.042 (0.744)	0.078 (1.250)

Table 5.6 Omit and Not Reached Information for Social Studies			
	Form A	Form B	Form C
Number of test takers	8,980	8,962	8,945
Number of items	50	50	50
Percent reaching all items	99.4	99.4	98.8
Percent Reaching 75% of items	99.9	99.9	99.8
Mean number of items omitted (standard deviation)	0.068 (0.575)	0.054 (0.385)	0.088 (0.717)
Mean number of items not reached (standard deviation)	0.037 (0.719)	0.053 (1.101)	0.063 (0.885)

Chapter 6: Differential Item Functioning

6.1 Overview

Differential item functioning (DIF) analyses were conducted separately for each subtest on the multiple-choice (MC) items. The DIF analyses were completed using *GENASYS*. DIF statistics are used to identify those items that identifiable groups of students (e.g. males and females) with the same underlying level of ability have different probabilities of answering correctly. If the item is more difficult for an identifiable subgroup when conditioned on ability, the item may be measuring something different from the intended construct. However, it is important to recognize that DIF-flagged items might be related to actual differences in relevant knowledge or skills (item impact) or statistical Type I error. As a result, DIF statistics are used to identify items that should be reviewed by ETS content experts from the DIF groups of interest to investigate the source and meaning of any apparent differences in item performance.

DIF analyses are conducted for designated comparison groups defined on the basis of gender and race/ethnicity, for any test on which the smaller of the two groups includes at least of 100 test takers and at least of 400 test takers are in both groups combined. Table 6.1 shows the DIF comparisons that were conducted on the HiSET subtests. The male and white groups are treated as the reference groups for gender and ethnicity, respectively; the female and other race and ethnic groups are considered the focal groups.

Table 6.1 DIF Comparisons

Comparison	Focal Group	Reference Group
Gender	Females	Males
Ethnicity	African-American	White
	Asian	White
	Hispanic	White
	Native American	White

Note. Sample sizes were insufficient to conduct DIF analyses for Pacific Islander test takers.

6.2 DIF Procedure

ETS uses the Mantel-Haenszel DIF detection method (Holland & Thayer, 1988) to compute a statistic called MH D-DIF[1]. This statistic indicates the difference between the focal and reference group performance on an item after conditioning on the total test score (the matching variable of ability). The difference is expressed on the "delta" scale, which is a transformation of the proportion correct, based on the inverse normal cumulative distribution function. Negative values imply that, conditional on the matching variable of ability, the focal group has a lower mean item score than the reference group — the focal group members' performance on the item was not as good as that of the reference group members with the same total score. In contrast, a positive value implies that, conditional on the matching variable, the reference group has a lower mean item score than the focal group — the focal group members' performance on the item was better than that of the reference group members with the same score.

6.3 DIF Flagging Criteria

The classification logic used for flagging items for DIF is based on a combination of absolute differences and significance testing. For items for which the statistical test indicates significant differences ($p < 0.05$), the effect size is used to determine the direction and magnitude of the DIF Based on the DIF statistics, items are classified into one of three categories and assigned values of A, B, or C. Category A items demonstrate negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large DIF.

[1] The formula for the estimate of constant odds ratio is

$$\hat{\alpha}_{MH} = \frac{\left(\sum_m \frac{R_{rm} W_{fm}}{N_m}\right)}{\left(\sum_m \frac{R_{fm} W_{rm}}{N_m}\right)},$$

where

R_{rm} = number in reference group at ability level m answering the item right,

W_{fm} = number in focal group at ability level m answering the item wrong,

R_{fm} = number in focal group at ability level m answering the item right,

W_{rm} = number in reference group at ability level m answering the item wrong,

N_m = total group at ability level m.

To facilitate the interpretation of MH results, the constant odds ratio is frequently transformed to the delta scale using the following formula (Holland & Thayer, 1988):

$$MH\ D\text{-}DIF = -2.35 \ln \hat{\alpha}_{MH}$$

Table 6.2 DIF Categories for Multiple-choice Items	
DIF Category	Criteria
A (negligible)	Absolute value of the MH D-DIF is not significantly different from zero, or is less than one.
B (slight to moderate)	1. Absolute value of the MH D-DIF is significantly different from zero but not from one, and is at least one; OR 2. Absolute value of the MH D-DIF is significantly different from one, but is less than 1.5. Positive values are classified as "B+" and negative values as "B-".
C (moderate to large)	Absolute value of the MH D-DIF is significantly different from one, and is at least 1.5. Positive values are classified as "C+" and negative values as "C-".

6.4 DIF Results

Tables 6.3 through 6.7 present the DIF results for the five HiSET subtests, combining the results across the three forms of each subtest. There were a few items flagged for C+ or C- DIF, across the comparison groups. The female/male DIF analyses resulted in between 0 and 2% of items (Reading and Social Studies) being flagged for C DIF, across tests. For race/ethnicity analyses, the largest number of items identified as having C DIF were within the Asian/White comparison group.

Table 6.3 Distribution of DIF Classifications for Reading						
Comparison Groups		DIF Categories				
		C+	B+	A	B-	C-
Female – Male	***N***	0	1	115	2	2
	%	0%	1%	96%	2%	2%
African American – White	***N***	0	3	114	3	0
	%	0%	3%	95%	3%	0%
Asian – White	***N***	3	14	86	13	4
	%	3%	12%	72%	11%	3%
Hispanic – White	***N***	2	5	105	6	2
	%	2%	4%	88%	5%	2%
Native American – White	***N***	0	4	110	6	0
	%	0%	3	92%	5%	0%

Note. Reading includes 120 MC items across three forms.

Table 6.4 Distribution of DIF Classifications for Writing

Comparison Groups		DIF Categories				
		C+	B+	A	B-	C-
Female – Male	***N***	0	3	142	4	1
	%	0%	2%	95%	3%	1%
African American – White	***N***	0	7	131	9	3
	%	0%	5%	87%	6%	2%
Asian – White	***N***	4	19	99	16	12
	%	3%	13%	66%	11%	8%
Hispanic – White	***N***	2	4	134	6	4
	%	1%	3%	89%	4%	3%
Native American – White	***N***	0	1	49	0	0
	%	0%	1%	33%	0%	0%

Note. Writing contains 150 MC items across three forms. The sample sizes were insufficient for many of the items to be analyzed for the Native American – White comparison group.

Table 6.5 Distribution of DIF Classifications for Mathematics

Comparison Groups		DIF Categories				
		C+	B+	A	B-	C-
Female – Male	***N***	0	1	139	9	1
	%	0%	1%	93%	6%	1%
African American – White	***N***	0	4	138	6	2
	%	0%	3%	92%	4%	1%
Asian – White	***N***	8	18	104	15	5
	%	5%	12%	69%	10%	3%
Hispanic – White	***N***	0	2	142	5	1
	%	0%	1%	95%	3%	1%
Native American – White	***N***	0	5	139	6	0
	%	0%	3%	93%	4%	0%

Note. Mathematics contains 150 MC items across three forms.

Table 6.6 Distribution of DIF Classifications for Science

Comparison Groups		DIF Categories				
		C+	B+	A	B-	C-
Female – Male	*N*	0	1	145	4	0
	%	0%	1%	97%	3%	0%
African American – White	*N*	0	0	143	7	0
	%	0%	0%	95%	5%	0%
Asian – White	*N*	1	18	118	9	4
	%	1%	12%	79%	6%	3%
Hispanic – White	*N*	0	1	144	5	0
	%	0%	1%	96%	3%	0%
Native American – White	*N*	0	4	44	2	0
	%	0%	3%	29%	1%	0%

Note. Science contains 150 MC items across three forms. The sample sizes were insufficient for many of the items to be analyzed for the Native American – White comparison group.

Table 6.7 Distribution of DIF Classifications for Social Studies

Comparison Groups		DIF Categories				
		C+	B+	A	B-	C-
Female – Male	*N*	0	6	134	7	3
	%	0%	4%	89%	5%	2%
African American – White	*N*	1	6	134	7	2
	%	1%	4%	89%	5%	1%
Asian – White	*N*	1	19	108	14	8
	%	1%	13%	72%	9%	5%
Hispanic – White	*N*	1	7	135	6	1
	%	1%	5%	90%	4%	1%
Native American – White	*N*	0	10	134	3	3
	%	0%	7%	89%	2%	2%

Note. Social Studies contains 150 MC items across three forms.

Chapter 7: Reliability

7.1 Overview

Reliability is the extent to which differences in test scores reflect true differences in the knowledge, ability, or skill being tested rather than fluctuations in performance due to chance. Thus, reliability is the consistency of the scores across conditions that can be assumed to differ at random, especially which form of the test the test taker is administered and which raters are assigned which constructed responses to score. In statistical terms, the variance in the distributions of test scores, a measure of the differences among individuals, is partly due to real differences in the knowledge, skill, or ability being tested ("true variance") and partly due to random differences in the measurement process ("error variance"). Reliability is an estimate of the proportion of the total variance that is true variance.

There are several different ways of estimating reliability. The type of reliability estimate reported in this technical report is an internal-consistency measure, which is derived from analysis of the consistency of the performance of individuals across items within a test. It is used because it serves as a good estimate of alternate forms reliability, but it does not take into account form-to-form variation due to lack of test form parallelism, nor is it responsive to day-to-day variation due to, for example, the test taker's state of health or the testing environment.

Reliability is enhanced when the component is maximized (e.g., internal consistency) or in other cases when it is minimized (errors). Reliability coefficients range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely individuals would be to obtain very similar scores upon repeated testing occasions, if the test takers do not change in their level of the knowledge or skills measured by the test. Sections 7.2 and 7.3 provide information regarding the estimation process and results.

Decision accuracy and decision consistency are also included in this report. Decision accuracy is the agreement between the classifications actually made and the classifications that would be made if the test scores were perfectly reliable. Decision consistency is the agreement between the classifications that would be made on two different forms of the test. Section 7.4 presents the results of decision accuracy and decision consistency analyses.

Interrater reliability is the reliability of the scoring process for the constructed response items, and is estimated from the agreement between individual raters (scorers). The interrater reliability coefficient answers the question, "How consistent would the scores of these test takers be over replication of scoring of the same responses by different scorers?" Section 7.5 provides information regarding calculation of interrater reliability and the corresponding results.

Standard error of measurement (*SEM*) quantifies the amount of error in the test scores. The *SEM* is the extent to which test takers' scores tend to differ from the scores they would receive if the average of the scores the person would have received on all the different forms of the test that could be made. There is a reliability coefficient and a corresponding *SEM* associated with each source, or combination of sources, of random variation that affect the scores. The formula for computing the *SEM* (see Formula 7-2) shows how the estimate of reliability and the *SEM* are related. A large *SEM* indicates that a test taker's score could have been quite different on a different form of the test. Observed scores with large *SEM*s pose a challenge to the valid interpretation of a single test score. Reliability and *SEM* estimates are calculated for each form of the five HiSET subtests.

7.2 Reliability and SEM Estimation

Coefficient alpha (Cronbach, 1951), which actually measures internal consistency, is commonly used to estimate alternative-forms reliability. Reliability estimates based on internal consistency measures are derived from analysis of the consistency of the performance of test takers across items within a test. These internal consistency measures serve as a good estimate of alternate forms reliability, but they are not responsive to day-to-day variation due to, for example, the test taker's state of health or the testing environment. Coefficient alpha is estimated by substituting sample estimates for the parameters in the formula:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_i^{2}}{\sigma_x^{2}}\right] \tag{7-1}$$

where n is the number of items, σ_i^2 is the variance of scores on the i-th item, and σ_x^2 is the variance of the total score (sum of scores on the individual items). Other things being equal, the more items a test includes, the higher the internal consistency reliability.

The formula for the standard error of measurement is:

$$\sigma_E = \sigma_x \sqrt{1 - \rho_{XX'}} \tag{7-2}$$

where σ_x is the standard deviation of the test total raw score, and p_{xx} is the reliability. The standard error is estimated by substitution of appropriate statistics for the parameters in equation 7-1.

7.3 Reliability Results for Total Group and Subgroups of Interest

Reliability estimates and corresponding *SEM*s of total test scores are presented, by form, in this section, for each subtest. The results are presented for all test takers combined, and for subgroups of interest. Tables 7.1, 7.3, 7.4, and 7.5 report the results for the MC-only tests (i.e., Reading, Mathematics, Science, and Social Studies). Tables 7.2a through 7.2c provide the results for Writing (MC and CR items), by form and by prompt. Overall, the reliability of the test forms containing only MC items ranged from 0.73 for Mathematics Form C to 0.87 for Reading Form C, with *SEM*s from 2.54 to 3.17. Reliability estimates for all forms and prompts of the Writing test, which included MC items and an essay, were between 0.71 and 0.72. *SEM*s were similar across forms and writing prompts (1.57 to 1.69).

Table 7.1 Test Reliability Estimates for Total Group and Subgroups, by Form: Reading

	Form A			Form B			Form C		
	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*
Total	9,045	0.85	2.57	8,910	0.83	2.54	8,619	0.87	2.58
Gender									
Male	4,782	0.85	2.56	4,659	0.84	2.53	4,551	0.87	2.58
Female	4,263	0.84	2.57	4,251	0.83	2.55	4,068	0.86	2.57
Race/Ethnicity									
American Indian	118	0.88	2.63	107	0.82	2.58	105	0.89	2.60
Asian	131	0.87	2.77	163	0.88	2.69	161	0.88	2.73
African American	1,570	0.80	2.78	1,596	0.81	2.69	1,448	0.82	2.74
White	4,499	0.82	2.44	4,397	0.82	2.45	4,313	0.85	2.47
Hispanic	1,484	0.82	2.64	1,405	0.81	2.60	1,393	0.84	2.69
Pacific Islander	*	*	*	*	*	*	*	*	*
Multiracial	308	0.80	2.49	292	0.80	2.49	303	0.86	2.48
Other/No Response	917	0.86	2.59	940	0.84	2.57	875	0.87	2.59

Note. Statistics not reported for sample size less than 25 ($N < 25$), denoted by '*'.

Table 7.2a Test Reliability Estimates for Total Group and Subgroups: Writing (MC and CR Items) – Form A

	Form A, Prompt 1			Form A, Prompt 2		
	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*
Total	3,053	0.71	1.64	3,019	0.71	1.61
Gender						
Male	1,612	0.72	1.65	1,593	0.71	1.64
Female	1,441	0.70	1.61	1,426	0.69	1.57
Race/Ethnicity						
American Indian	36	0.73	1.89	47	0.71	1.75
Asian	56	0.72	1.93	48	0.72	1.98
African American	468	0.69	1.66	441	0.70	1.60
White	1,620	0.71	1.57	1,609	0.71	1.57
Hispanic	450	0.68	1.59	463	0.68	1.54
Pacific Islander	*	*	*	*	*	*
Multiracial	98	0.74	1.43	105	0.69	1.62
Other/No Response	320	0.71	1.64	302	0.71	1.62

Note. Statistics not reported for sample size less than 25 ($N < 25$), denoted by '*'.

Table 7.2b Test Reliability Estimates for Total Group and Subgroups: Writing (MC and CR Items) - Form B

	Form B, Prompt 1			Form B, Prompt 2			Form B, Prompt 3		
	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*
Total	2,952	0.71	1.69	3,004	0.71	1.66	3,044	0.71	1.69
Gender									
Male	1,540	0.71	1.73	1,641	0.71	1.70	1,644	0.71	1.71
Female	1,412	0.71	1.64	1,363	0.71	1.61	1,400	0.70	1.65
Race/Ethnicity									
American Indian	39	0.68	1.53	41	0.74	1.84	34	0.74	1.92
Asian	39	0.71	2.07	48	0.74	2.12	40	0.76	2.00
African American	472	0.65	1.65	476	0.67	1.58	479	0.69	1.66
White	1,537	0.72	1.64	1,582	0.72	1.60	1,571	0.71	1.65
Hispanic	445	0.69	1.68	444	0.69	1.68	503	0.67	1.63
Pacific Islander	*	*	*	*	*	*	*	*	*
Multiracial	90	0.73	1.62	99	0.73	1.61	118	0.73	1.69
Other/No Response	324	0.73	1.69	308	0.70	1.69	291	0.70	1.66

Note. Statistics not reported for sample size less than 25 ($N < 25$), denoted by '*'.

Table 7.2c Test Reliability Estimates for Total Group and Subgroups: Writing (MC and CR Items) - Form C

	Form C, Prompt 1			Form C, Prompt 2			Form C, Prompt 3		
	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*
Total	2,956	0.72	1.66	2,989	0.72	1.63	3,004	0.72	1.57
Gender									
Male	1,552	0.72	1.65	1,631	0.72	1.63	1,629	0.73	1.58
Female	1,404	0.72	1.65	1,358	0.72	1.62	1,375	0.72	1.56
Race/Ethnicity									
American Indian	27	0.56	1.63	35	0.73	1.65	33	0.71	1.63
Asian	39	0.75	1.94	58	0.71	1.79	32	0.73	1.64
African American	498	0.70	1.60	493	0.69	1.55	490	0.69	1.50
White	1,523	0.72	1.60	1,556	0.73	1.61	1,583	0.73	1.55
Hispanic	432	0.68	1.58	447	0.69	1.55	461	0.70	1.48
Pacific Islander	*	*	*	*	*	*	*	*	*
Multiracial	135	0.73	1.59	107	0.72	1.55	109	0.69	1.48
Other/No Response	297	0.73	1.69	287	0.72	1.61	293	0.72	1.57

Note. Statistics not reported for sample size less than 25 ($N < 25$), denoted by '*'.

Table 7.3 Test Reliability Estimates for Total Group and Subgroups, by Form: Mathematics

	Form A			Form B			Form C		
	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*
Total	10,149	0.75	2.96	9,937	0.76	2.90	10,316	0.73	2.85
Gender									
Male	5,182	0.77	2.96	5,061	0.77	2.92	5,239	0.74	2.86
Female	4,967	0.72	2.94	4,876	0.72	2.88	5,077	0.69	2.84
Race/Ethnicity									
American Indian	133	0.68	2.90	107	0.74	2.89	133	0.66	2.87
Asian	139	0.90	2.97	169	0.89	2.94	147	0.87	2.88
African American	1,967	0.61	2.93	1,854	0.59	2.82	1,916	0.56	2.82
White	4,846	0.77	2.95	4,865	0.77	2.92	5,109	0.74	2.85
Hispanic	1,620	0.68	2.96	1,542	0.67	2.89	1,636	0.63	2.84
Pacific Islander	26	0.68	2.95	*	*	*	*	*	*
Multiracial	367	0.77	2.95	338	0.79	2.96	347	0.75	2.85
Other/No Response	1,051	0.75	2.97	1,042	0.75	2.90	1,008	0.73	2.86

Note. Statistics not reported for sample size less than 25 (*N* < 25), denoted by '*'.

Table 7.4 Test Reliability Estimates for Total Group and Subgroups, by Form: Science

	Form A			Form B			Form C		
	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*
Total	8,349	0.84	3.03	8,414	0.85	3.07	8,235	0.86	3.05
Gender									
Male	4,405	0.84	3.00	4,376	0.86	3.03	4,346	0.86	3.01
Female	3,944	0.83	3.05	4,038	0.83	3.10	3,889	0.85	3.09
Race/Ethnicity									
American Indian	114	0.85	3.03	92	0.82	3.13	96	0.86	3.12
Asian	124	0.87	3.06	134	0.86	3.15	113	0.90	3.05
African American	1,429	0.77	3.13	1,422	0.77	3.20	1,426	0.79	3.19
White	4,161	0.82	2.95	4,326	0.85	2.98	4,213	0.84	2.97
Hispanic	1,333	0.80	3.10	1,303	0.80	3.13	1,244	0.83	3.13
Pacific Islander	*	*	*	*	*	*	*	*	*
Multiracial	313	0.82	2.97	271	0.82	3.02	292	0.84	2.99
Other/No Response	864	0.84	3.06	847	0.85	3.08	835	0.85	3.07

Note. Statistics not reported for sample size less than 25 ($N < 25$), denoted by '*'.

Table 7.5 Test Reliability Estimates for Total Group and Subgroups, by Form: Social Studies

	Form A			Form B			Form C		
	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*	Number of Test Takers	Reliability Estimate	*SEM*
Total	8,980	0.82	3.11	8,962	0.85	3.08	8,945	0.83	3.17
Gender									
Male	4,589	0.83	3.08	4,639	0.85	3.04	4,639	0.84	3.14
Female	4,391	0.81	3.13	4,323	0.83	3.11	4,306	0.81	3.19
Race/Ethnicity									
American Indian	101	0.81	3.13	121	0.85	3.09	118	0.80	3.23
Asian	146	0.86	3.09	141	0.86	3.17	130	0.84	3.21
African American	1,592	0.74	3.18	1,606	0.77	3.19	1,554	0.75	3.24
White	4,482	0.82	3.05	4,403	0.84	3.01	4,481	0.83	3.11
Hispanic	1,452	0.77	3.17	1,447	0.81	3.15	1,455	0.80	3.21
Pacific Islander	*	*	*	*	*	*	*	*	*
Multiracial	312	0.79	3.10	307	0.83	2.99	290	0.82	3.14
Other/No Response	882	0.82	3.12	924	0.85	3.09	896	0.84	3.18

Note. Statistics not reported for sample size less than 25 ($N < 25$), denoted by '*'.

7.4 Reliability of Classification

The reliability of the classifications (i.e., pass/fail high school equivalency; pass/fail college and career readiness) for the test takers was calculated using the computer program *RELCLASS* (ETS proprietary software), which operationalizes a statistical method developed by Livingston and Lewis (1995). This method uses information from the administration of one test form (i.e., distribution of scores, the minimum and maximum possible scores, the cut points used for classification, and the reliability coefficient) to estimate two kinds of statistics, "decision accuracy" and "decision consistency." Decision accuracy refers to the extent to which the classifications of test takers based on their scores on the test form agree with the classifications that would be made if each person's average score over all possible forms of the test could be known. Decision consistency refers to the agreement between the classifications based on two non-overlapping, equally difficult forms of the test.

Note that in all cases the decision accuracy indices are somewhat larger than the decision consistency indices. For decision accuracy, only the observed-score classification is affected by random variation; the true-score classification is not affected by random variation. For decision consistency, each of the two classifications is based on a score that is affected by random variation (Livingston & Lewis, 1995).

Tables 7.6 through 7.10 provide information regarding the accuracy and consistency of the two classifications made on the basis of HiSET scores: *High School Equivalency Cut Point* (i.e., did the test taker demonstrate high school equivalency on each subtest) and *College and Career Cut Point* (i.e., was the test taker classified as meeting College and Career Readiness). These results are presented by form for each test. The decision accuracy indices for the *High School Equivalency Cut Point* ranged from 0.83 for all three forms of Mathematics to 0.96 for Writing Form A using Prompt 2; while the corresponding decision consistency indices ranged from 0.76 for Mathematics Form C to 0.94 for Writing Form A using Prompt 2. Decision accuracy values for the *College and Career Cut Point* ranged from 0.87 for Writing Form A using Prompt 1 to 0.94 for all three forms of Mathematics. Parallel decision consistency values ranged from 0.82 for Writing Form A using Prompt 1 to 0.91 for all three forms of Mathematics.

Table 7.6 Classification Consistency and Accuracy for Reading

Form	*N*	Accuracy			Consistency			HS Equivalency Cut Point		College and Career Cut Point	
		Overall	False Positive	False Negative	Overall	False Positive	False Negative	Cut Point Accuracy	Cut Point Consistency	Cut Point Accuracy	Cut Point Consistency
Form A	9,045	0.82	0.08	0.10	0.75	0.12	0.13	0.93	0.91	0.89	0.84
Form B	8,910	0.81	0.08	0.12	0.73	0.12	0.15	0.92	0.89	0.89	0.84
Form C	8,619	0.81	0.06	0.13	0.74	0.10	0.16	0.92	0.89	0.89	0.84

Note. For both Accuracy and Consistency, False Positive refers to test takers who were estimated to be incorrectly classified as achieving High School Equivalency on the test or being college and career ready; False Negative refers to test takers who were estimated to be incorrectly classified as **not** achieving High School Equivalency on the test or **not** being college and career ready.

Table 7.7 Classification Consistency and Accuracy for Writing

Form	*N*	Accuracy			Consistency			HS Equivalency Cut Point		College and Career Cut Point	
		Overall	False Positive	False Negative	Overall	False Positive	False Negative	Cut Point Accuracy	Cut Point Consistency	Cut Point Accuracy	Cut Point Consistency
Form A, Prompt 1	3,092	0.83	0.08	0.09	0.76	0.12	0.13	0.95	0.93	0.87	0.82
Form A, Prompt 2	3,050	0.83	0.08	0.09	0.77	0.11	0.12	0.96	0.94	0.88	0.83
Form B, Prompt 1	2,978	0.83	0.09	0.08	0.76	0.12	0.12	0.95	0.93	0.88	0.83
Form B, Prompt 2	3,034	0.83	0.09	0.08	0.77	0.12	0.12	0.95	0.92	0.89	0.84
From B, Prompt 3	3,064	0.83	0.08	0.08	0.77	0.12	0.12	0.95	0.92	0.89	0.84
Form C, Prompt 1	2,974	0.84	0.08	0.08	0.78	0.11	0.11	0.94	0.92	0.90	0.86
Form C, Prompt 2	3,013	0.84	0.07	0.08	0.78	0.11	0.11	0.94	0.92	0.90	0.86
From C, Prompt 3	3,018	0.85	0.07	0.08	0.79	0.10	0.11	0.95	0.93	0.90	0.86

Note. For both Accuracy and Consistency, False Positive refers to test takers who were estimated to be incorrectly classified as achieving High School Equivalency the test <u>or</u> being college and career ready; False Negative refers to test takers who were estimated to be incorrectly classified as **not** achieving High School Equivalency the test <u>or</u> **not** being college and career ready.

Table 7.8 Classification Consistency and Accuracy for Mathematics

Form	*N*	Accuracy			Consistency			HS Equivalency Cut Point		College and Career Cut Point	
		Overall	False Positive	False Negative	Overall	False Positive	False Negative	Cut Point Accuracy	Cut Point Consistency	Cut Point Accuracy	Cut Point Consistency
Form A	10,149	0.77	0.11	0.11	0.69	0.15	0.16	0.83	0.77	0.94	0.91
Form B	9,937	0.77	0.12	0.11	0.68	0.15	0.16	0.83	0.77	0.94	0.91
Form C	10,316	0.76	0.13	0.11	0.67	0.17	0.16	0.83	0.76	0.94	0.91

Note. For both Accuracy and Consistency, False Positive refers to test takers who were estimated to be incorrectly classified as achieving High School Equivalency the test or being college and career ready; False Negative refers to test takers who were estimated to be incorrectly classified as **not** achieving High School Equivalency the test or **not** being college and career ready.

Table 7.9 Classification Consistency and Accuracy for Science

Form	*N*	Accuracy			Consistency			HS Equivalency Cut Point		College and Career Cut Point	
		Overall	False Positive	False Negative	Overall	False Positive	False Negative	Cut Point Accuracy	Cut Point Consistency	Cut Point Accuracy	Cut Point Consistency
Form A	8,349	0.83	0.08	0.09	0.76	0.12	0.12	0.94	0.91	0.89	0.84
Form B	8,414	0.83	0.08	0.09	0.76	0.11	0.12	0.95	0.93	0.88	0.83
Form C	8,235	0.83	0.08	0.09	0.76	0.12	0.12	0.95	0.93	0.88	0.83

Note. For both Accuracy and Consistency, False Positive refers to test takers who were estimated to be incorrectly classified as achieving High School Equivalency the test or being college and career ready; False Negative refers to test takers who were estimated to be incorrectly classified as **not** achieving High School Equivalency the test or **not** being college and career ready.

Table 7.10 Classification Consistency and Accuracy for Social Studies

Form	*N*	Accuracy			Consistency			HS Equivalency Cut Point		College and Career Cut Point	
		Overall	False Positive	False Negative	Overall	False Positive	False Negative	Cut Point Accuracy	Cut Point Consistency	Cut Point Accuracy	Cut Point Consistency
Form A	8,980	0.79	0.12	0.10	0.70	0.16	0.14	0.89	0.85	0.89	0.85
Form B	8,962	0.79	0.10	0.11	0.72	0.14	0.14	0.89	0.84	0.91	0.87
Form C	8,945	0.79	0.09	0.12	0.71	0.14	0.16	0.88	0.84	0.90	0.86

Note. For both Accuracy and Consistency, False Positive refers to test takers who were estimated to be incorrectly classified as achieving High School Equivalency the test or being college and career ready; False Negative refers to test takers who were estimated to be incorrectly classified as **not** achieving High School Equivalency the test or **not** being college and career ready.

7.5 Interrater Agreement

Rater agreement or consistency is critical for valid test score interpretation of assessments requiring human raters to rate the essay responses. When two trained raters independently assign the same score (or rating) to a test taker's item response, there is evidence that the scoring standard is being applied consistently. Double scoring substantially increases the reliability of the scoring process. Double scoring is used to monitor and evaluate the accuracy of rating, 100% of the responses are rated twice. Interrater reliability is evaluated empirically in three different statistics: a) Percentage agreement between two raters, b) Intraclass correlation, and c) Weighted kappa coefficient.

a) Percentage of exact score agreement becomes a more stringent criterion as the number of item score points in the rating scale increases. The fewer the item score points, the fewer degrees of freedom on which two raters can vary (i.e., the fewer ratings the two raters can make differently), and the higher the percentage of agreement is likely to be. For the essay component of the Writing test, the rating scale ranges from 0 to 6. The percentage of exact agreement, the percentage of disagreement by 1 scale score point, and the percentage of disagreement by 2 or more scale score points were considered when evaluating the differences between ratings on each essay prompt.

b) The intraclass correlation, r_{IC} is the proportion of variance that is consistent between raters scoring the same essays. The range of intraclass correlation is from 0.0 to 1.0, with 1.0 indicating perfect agreement between the first and second raters. Suppose that N is the number of responses that are scored twice, X_{n1} and X_{n2} are the two scores of response n, n = 1, 2, … N):

$$r_{IC} = \frac{\frac{1}{N-1}\sum_{n=1}^{N}\left[\left(\overline{X}_{n.} - \overline{X}_{..}\right)^2\right]}{\frac{1}{2(N-1)}\sum_{n=1}^{N}\left[\left(X_{n1} - \overline{X}_{..}\right)^2 + \left(X_{n2} - \overline{X}_{..}\right)^2\right]}, \quad (7\text{-}3)$$

where

$$\overline{X}_{n.} = \left(X_{n1} + X_{n2}\right)/2, \quad (7\text{-}4)$$

and

$$\overline{X}_{..} = \frac{1}{N}\sum_{n=1}^{N}\left(X_{n1} + X_{n2}\right)/2. \quad (7\text{-}5)$$

While intraclass correlations were calculated for each of the 8 Writing essay prompts (two prompts for Form A; three prompts each for Forms B and C, these statistics are not presented in this technical report. When the distribution of scores on a given prompt is the same for the first and second raters, the intraclass correlation will be equal to weighted kappa (Fleiss & Cohen, 1973).

c) The quadratic weighted kappa coefficient was selected because unweighted kappa does not take into account the degree of disagreement between raters. The quadratic weighted kappa is a generalization of the simple kappa coefficient using weights to quantify the relative difference between categories. The range of quadratic weighted kappa coefficients is from 0.0 to 1.0, with perfect agreement indicated by 1.0.

For an item with m categories, one can construct an m x m rating table with scores provided by two raters, A and B. Suppose m is the maximum obtainable rating for each item,

- n_{ij} is the number of responses for which rater A's rating = i and rater B's rating = j,
- n_{i+} is the number of responses for which Rater A = i,
- n_{+j} is the number of responses for which Rater B = j,
- and n_{++} is the number of all responses from raters A and B. The quadratic weighted kappa coefficient is defined as:

$$k_{ij} = \frac{\left(\sum_{i=0}^{m}\sum_{j=0}^{m} w_{ij}\frac{n_{ij}}{n_{++}}\right) - \left(\sum_{i=0}^{m}\sum_{j=0}^{m} w_{ij}\frac{n_{i+}n_{+j}}{n_{++}^2}\right)}{1-\left(\sum_{i=0}^{m}\sum_{j=0}^{m} w_{ij}\frac{n_{i+}n_{+j}}{n_{++}^2}\right)}, \tag{7-6}$$

where

$$w_{ij} = 1 - \frac{(i-j)^2}{m^2}. \tag{7-7}$$

Table 7.11 presents the mean scale scores, agreement rates, and quadratic weighted kappa for the HiSET essays. There is one essay on each form of the Writing test and each essay is scored by two raters. Each essay has a maximum possible scale score of 6. Exact agreement ranged from 58% (Form B, Prompt 1) to 61% (Form C, Prompt 3). Form C, Prompt 3 had the lowest percentage of difference by one scale score point (37%), while Form B, Prompt 2 had the highest percentage of one scale score point difference (40%). Interrater differences of two or more scale score points were relatively low ranging from 2% to 3%. The weighted kappa coefficients were moderately high, ranging from 0.71 (Form C, Prompt 3) to 0.76 (Form C, Prompt 1).

Table 7.11 Interrater Agreement for Writing Essay

Prompt ID	Mean Score	Absolute Difference (Percentage)			Weighted Kappa
		No Difference	1 Point	2 or more Points	
Form A, Prompt 1	3.11	58	39	2	0.73
Form A, Prompt 2	3.19	60	38	3	0.72
Form B, Prompt 1	3.29	58	40	3	0.72
Form B, Prompt 2	3.17	58	40	2	0.72
Form B, Prompt 3	3.32	59	39	2	0.73
Form C, Prompt 1	3.26	60	38	2	0.76
Form C, Prompt 2	3.25	59	39	2	0.72
Form C, Prompt 3	3.29	61	37	2	0.71

Chapter 8: Validity

8.1 Overview

The Standards for Educational and Psychological Testing, issued jointly by the American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014) states the following:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations (p. 11).

The purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for particular uses. Test validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the lifetime of an assessment. Every aspect of an assessment provides evidence in support of its validity (or evidence of lack of validity), including design, content specifications, item development, and psychometric characteristics.

Test takers' scores on the HiSET exam are intended to reflect their level of knowledge and skills for each of the five HiSET subtests. The scores are used to classify test takers in terms of their level of proficiency with regard to high school equivalency and college and career readiness. A standard setting meeting was held to establish the cut scores for high school equivalency and college and career readiness (Tannenbaum & Reese, 2014). Although there are minimum cut scores, individual states may choose to raise the cut scores for awarding a high school equivalency certificate or for being identified as college and career ready.

8.2 Validity Evidence Based on Test Content

The HiSET development process began with a review of the CCRS (Pimentel, 2013; https://lincs.ed.gov/publications/pdf/CCRStandardsAdultEd.pdf) that describes the skills and knowledge that adults and youth who have not graduated from high school should acquire to successfully be prepared to enter a job, a training program, or an entry-level, credit-bearing postsecondary course. The test development process, including the content framework is described in Section 2.1 of this Technical Report.

8.2.1 Fairness

Concern for fairness and the elimination of bias from the assessment is a guiding principle throughout design and development. In particular, the HiSET assessment is built with careful attention to content-related sources of test bias. Procedures address this source of bias through the following:

- Thorough examination of content and performance standards for the selection of the appropriate content. The bias reviews were conducted by Iowa Testing Program (ITP) that developed the HiSET items. A summary of the fairness activities is provided in Section 2.1.2.
- Engagement of panels of experts in the review of the test specifications, items, forms, and the essay scoring rubrics. A brief summary of the content reviews conducted by ITP is provided in Section 2.1.1.

- Alignment of items to the College and Career Readiness Standards (see Section 8.2.2),
- Statistical procedures for identifying items on these tests that function differently across various groups of test takers (see Chapter 6 for a description of the DIF analyses and the results), and
- Careful selection of a national sample of test takers to respond to the assessment. The norming samples were selected to represent the diverse characteristics of high school seniors based on gender, ethnicity, district size, region of the country, and socioeconomic characteristics of the school (see the Norms and Norming Samples section in Chapter 3 of the HiSET technical manual, Educational Testing Service, 2014b).

8.2.2 Alignment to the College and Career Readiness Standards for Adult Education

ETS contracted WestEd to conduct an independent third-party alignment study of the HiSET items. In an alignment study, the degree to which the items represent the content of standards is examined (Webb, 1999). WestEd evaluated alignment of the Reading and Mathematics HiSET items to the College and Career Readiness Standards (CCRS) which were released by the U.S. Department of Education in 2013 (https://lincs.ed.gov/publications/pdf/CCRStandardsAdultEd.pdf). Researchers at WestEd evaluated each item to the CCRS using a modified Webb-based procedure (Webb, 1999, 2002, 2007). The level of match between the HiSET item and the CCRS was categorized as:

- Strong Alignment: substantial or foundational overlap between the CCRS and the item, additionally the item measured the same central idea, fundamental skill, or core concept as the CCRS,
- Partial Alignment: some overlap between the CCRS and the item but the relationship is weaker, and
- No Alignment: no overlap between the CCRS and the item.

The WestEd researchers found that 87% of the 120 Reading items and 88% of the 150 Mathematics items had either a Strong or Partial Alignment with the CCRS. Additionally, the WestEd researchers evaluated the education level of the items ranging from Beginning adult basic education literacy (kindergarten and Grade 1 levels) to Low/High adult secondary education (Grades 9 through Grade 12). Out of the 120 ELA items, 75% of the items were classified as low intermediate basic education and above — targeting knowledge and skills at the Grades 4 through 12 levels. For Mathematics, the majority of item (79%) were classified as targeting the knowledge and skills of Grades 6 through 12. Based on the findings, the researchers at WestEd concluded that there are high rates of alignment (strong and partial) between the HiSET items and the CCRS.

8.3 Construct Validity in Support of Content Structure

The general constructs underlying the HiSET program were investigated using exploratory factor analysis techniques (Browne, 1979; Schreiber, Nora, Stage, Barlow, & King, 2006; Tucker, 1958). The identified factors clearly reflect the test composition and are consistent with the emphasis found in high school curricula (ETS, 2014b). The first factor could be identified as a "literacy" factor, the second factor is a "numeracy" factor, and the third factor is described as an "analysis of information" factor. The Reading test contributes the most to the interpretation of the first factor, with some additional contribution from the Writing, Social Studies, and Science tests. The Mathematics test loads heavily on the second factor with some limited contribution from Science. The Science and Social Studies tests are most clearly associated with the third factor.

8.3.1 Validity Evidence Based on Internal Test Structure

The internal structure of the HiSET exam is assessed in relation to the degree to which these tests meet the requirements of the statistical models used to estimate item parameters and test taker scores. Confirmatory factor analyses (CFAs) was conducted to validate the underlying domain structure of each HiSET subtest. CFA is a useful statistical methodology as it can be used to evaluate whether performance on items in each subtest reflects a single underlying dimension. The findings from this type of analysis provide evidence as to whether the unidimensional model-based IRT used to calibrate the HiSET items is appropriate. Additionally, when reporting a single scale score for a subtest an assumption is made that all the items on the test measure the same single underlying dimension. Therefore, CFAs are also useful for supporting the reporting of a single scale score for each HiSET subtest.

8.3.2 Confirmatory Factor Analyses of the Tests

Confirmatory factor analyses (CFAs) were performed to evaluate the dimensionality of the data because the 3PL IRT model assumes unidimensionality of the data. To evaluate the dimensionality of the HiSET exam, CFAs were conducted using test data from one of the three forms within each subtest. The form chosen for analysis had the largest number of test takers. Both one-factor and multi-factor models were investigated. The multi-factor models were identified by the subscore structure for each subtest, as determined by content specialists.

Mplus (L. K. Muthén & Muthén, 1998–2012) was used to calculate matrices of polychoric correlations between the items included in each analysis. Mplus was also used to fit specified factor models to the data. In the analysis, the input polychoric correlation matrix was used to estimate the factor loadings on the indicators (items).

Parameter estimation was accomplished using a weighted least-squares method with mean and variance adjustment (B. Muthén, DuToit, & Spisic, 1997). This method leads to a consistent estimator of the model parameters, and provides standard errors that are robust under model misspecification. For ordinal data, such as the scores for the written essay, weighted least squares estimation offers an alternative to full-information maximum likelihood techniques. The latter becomes too computationally demanding for models with more than a few dimensions. Model fit can be assessed through the use of a scaled chi-square statistic. However, the degrees of freedom for the reference distribution of this statistic cannot be computed in the standard way. The correct degrees of freedom are in part determined by the data, and different degrees of freedom may be obtained when applying the same model to different data (B. Muthén, 1998–2004, p. 19–20).

Overall model fit for each CFA model within each subtest was examined using several fit indices. The Tucker-Lewis Index (TLI) compares the chi-square for the hypothesized model with that of the null or "independence" model, in which all correlations or covariances are zero. TLI values range from zero to 1.0, and, as a general rule of thumb, values greater than 0.90 signify acceptable fit (Hu & Bentler, 1999). The comparative fit index (CFI) and root mean square error of approximation (RMSEA) index both are based on noncentrality parameters. The CFI compares the covariance matrix predicted by the model with the observed covariance matrix, and the covariance matrix of the null model with the observed covariance matrix. A CFI value greater than 0.90 indicates acceptable model fit (Hu & Bentler, 1999). The RMSEA assesses the error in the hypothesized model predictions; values less than or equal to 0.06 indicate good fit (Hu & Bentler, 1999).

Table 8.1 shows the results of the one-factor CFAs. The TLI, CFI, and RMSEA fit statistics indicate that the one-factor solutions provide acceptable fit for Reading, Writing, Science, and Social Studies. The one-factor model fit Mathematics based on the RMSEA.

Multi-factor CFAs[3] with items loading on different content categories or subscores did not provide improved model fit for any of the subtests when compared to the results of the one-factor models. There were high estimated correlations among latent factors for all subtests. Estimated correlations between the latent factors were greater than 1.0 for both Reading and Writing, resulting in latent variable covariance matrices being non-positive definite. These linear dependencies could be resolved by combining some factors for Reading and for Writing, but the estimated correlations among the reduced number of latent factors were still very high, indicating that multi-factor structures for Reading and for Writing are not well supported.

These findings provide evidence that a single dimension or factor exists for each of the five HiSET subtests. This is a positive outcome, given that IRT models assume unidimensionality, and the 3PL IRT model was used as the equating method for the HiSET exam.

Table 8.1 Confirmatory Factor Analyses Fit Statistics: One-factor Model

Content	Form	# of Factors	# of Items	*N*	TLI	CFI	RMSEA
Reading	B	1	40	8,910	**0.978**	**0.979**	**0.015**
Writing	A	1	51	3,092	**0.941**	**0.943**	**0.018**
Mathematics	C	1	50	10,316	0.797	0.805	**0.023**
Science	B	1	50	8,414	**0.908**	**0.912**	**0.031**
Social Studies	B	1	50	8,962	**0.914**	**0.918**	**0.027**

Note. Data from Form A, Prompt 1 was used for the Writing CFA. Table entries that meet or exceed the criterion for acceptable fit are in bold.

8.4 Correlations between HiSET Subtests

The relationship of the scores between subtests was evaluated using correlational analyses. The results are presented in Table 8.2. The degree to which the subtest scores correlate provides evidence that the tests measure different constructs. The correlations are consistent with expectations in that scores from the five subtests are only moderately associated, with correlations ranging from 0.50 to 0.74. These intercorrelations are lower than the reliability estimates reported in Tables 7.1 through 7.5. Therefore, the items within each subtest are more strongly correlated with each other, than the items across subtests (i.e., the Reading items are more correlated with other Reading items than with items from the other four HiSET subtests). For example, the estimates of reliability for Reading ranged from 0.83 to 0.87 which are all higher than the correlations between Reading and the other four HiSET tests.

[3] The results of the multi-factor analyses are not provided in this technical report but are available upon request.

Table 8.2 Correlations between Subtests					
	Reading	Writing	Mathematics	Science	Social Studies
Reading	1.00				
Writing	0.69	1.00			
Mathematics	0.50	0.54	1.00		
Science	0.71	0.66	0.63	1.00	
Social Studies	0.74	0.66	0.57	0.74	1.00

Note. The reliability estimates, as reported in Tables 7.1 through 7.5, are: Reading 0.83 to 0.87, Writing: 0.71 to 0.72, Math: 0.73 to 0.75, Science 0.84 to 0.86, and Social Studies 0.82 to 0.85.

8.5 Validity Evidence from the Special Studies

Research studies will be conducted to provide additional validity evidence in support of the HiSET program's intended interpretations and uses of test scores. Two of the special studies that ETS is currently planning to conduct are:

- An investigation into the relationship between HiSET subtest scores and high school performance, and
- An investigation of the educational and employment outcomes for test takers who successfully complete the HiSET battery.

The following paragraphs briefly describe these planned studies. (Note, that these descriptions reflect what was proposed. The specifics of each study's actual implementation may vary somewhat from what was proposed).

To support the claim that the HiSET exam is a measure of high school equivalency and a measure of college readiness, researchers at ETS will conduct a study to evaluate whether passing the HiSET battery is equivalent to successful performance in high school. The researchers will look at the relationships between scores on the HiSET subtests and measures of high school performance, such as high school grade point average or scores on high school exit exams. For this study, it is expected that over 3,000 high school seniors will each take two of the HiSET subtests. The 3,000 high school seniors will be sampled from several U.S. states from over 140 high schools, representing urban, suburban, and rural high schools. Additionally demographic information (e.g., gender and ethnicity) will be considered so that a variety of ethnic groups will be represented. Additionally, the ETS researchers plan on evaluating the relationship between performance on the HiSET subtests and college readiness assessments.

Other researchers at ETS plan to investigate the educational and employment outcomes for the test takers who successfully complete the HiSET battery and have received high school equivalency credentials. The researchers will do a time series study tracking the test takers to determine if completing the HiSET battery results in new educational or employment opportunities. The researchers will track approximately 2,500 successful HiSET test takers over a five year period. Each year the researchers will ask the participants to provide information on their current educational or employment status. Additionally, factors impacting test takers' employment and educational status will be collected and evaluated.

These two special studies are in the early stages of implementation. Therefore, the results, when available, will be summarized in future technical reports.

Chapter 9: Establishment and Maintenance of Score Scales

9.1 The HiSET Score Scale

HiSET scores for each subtest (Reading, Writing, Mathematics, Science, and Social Studies) are reported on a 1–20 scale at integer values. The most important aspect of a score scale is not the selection of the score values themselves (the integers 1 through 20 in this case) but rather how test performance levels are associated to the reported values.

Performance on each HiSET subtest is most directly measured by the numbers of questions answered correctly. The HiSET Mathematics, Science, Social Studies and Writing subtests each contain 50 items, while Reading contains 40 items. As such, number-correct scores range from 0 to 50 for most tests, while Reading number-correct scores range from 0 to 40. Each number-correct score on each test maps to a corresponding value on the 1–20 reported score scale. This chapter describes both how the mapping between number-correct and reported scores was established for the initial set of HiSET forms that were administered in 2014 and how the mapping is determined for each new HiSET form in such a way that reported scores represent comparable performance levels regardless of the test form on which they were achieved.

As will be described more fully below, scoring and scaling procedures for Writing differ from those applied to the other four subtests. These differences result from the Writing test being composed of two distinct sections. The first is a section of 50 objectively scored items while the second is a single writing sample or essay. Both sections are first scored and scaled independently. Number-correct scores on the 50-item objectively-scored MC section are converted to a 1–14 scale (rather than a 1–20 scale). The final essay score is reported using a 0–6 scale. The total Writing scores are then computed as the sum of the two components (the 50-item MC section and the 6-point essay section), placing them on the standard 1–20 scale.

9.2 Establishing the Initial HiSET Reported Score Scale

The choice of the integers 1 through 20 to convey HiSET performance levels was made considering a few basic principles. First, the reported scores should not be easily confused with either scores on other tests or with other common metrics for performance. For example, reporting HiSET scores on the ranges 1–36 or 200–800 would have risked confusion with the ACT or SAT college entrance exam subscores, respectively. Similarly, reporting scores on a 0–100 scale might have risked confusion with percentage correct or percentile rank.

A second principle in choosing a score scale is that the scale should not imply that performance is measured at a finer grain than the test truly allows. For example, mapping the 0–50 number-correct scores to a 0–200 reported score scale might imply that there are more specific levels of performance than the test in fact permits. Because only 51 unique performance levels could be measured by each test form most of the available reported score values would go unused.

A third principle applies to tests like the HiSET exam that report scores across multiple subtests. This principle holds that different tests in the same battery should be similarly scaled. For example, HiSET scores for the subtests are reported on the same 1–20 scale. However, greater similarity is desirable to avoid a common sort of score misinterpretation, best illustrated by example. Consider a test taker who scored 13 on the

HiSET Mathematics test and 15 on Reading. It might be reasonably concluded that this test taker performed relatively better in Reading than in Mathematics. However, suppose that average scores for all test takers were 10 in Mathematics and 16 in Reading. The test taker then performed above average in Mathematics and below average in Reading, opposite the conclusion suggested by the scores themselves. A typical way of complying with this principle is to set the score scales for each subtest so that the average scaled score for all test takers is the same in all five subtests.

A fourth principle is closely related to the third, above. Because the HiSET program imposes a passing threshold on each test, it is most convenient if these thresholds are identical across tests. This means that the minimal passing performance within each subtest is associated with the same scaled score. The passing scaled score was selected as 8 in each subtest, but the number of correct responses corresponding with the scaled score of 8 varies across test forms and subtests. The process of determining the number of correct responses associated with a passing threshold is called *standard setting* and is described fully in the HiSET standard setting report (Tannenbaum & Reese, 2014).

The standard setting process gathered panels of adult educators and subject matter experts who examined the items on selected HiSET test forms and judged the number that a candidate minimally qualified to be deemed as passing would answer correctly. These judgments were successively refined and then averaged across panelists to determine the passing standard in terms of number-correct scores on each of the selected forms. The form selected for examination in each subtest was then designated as the *base* form and the number of correct responses needed for passing was associated with a scaled score of 8. This correspondence is summarized in Table 9.1. Summing these two components produces the usual total score threshold of eight. The standard setting for Writing addressed the objectively-scored and essay components of the test independently. The passing threshold on the objectively-scored section was mapped to a scaled score of six (rather than eight) while the passing essay performance was mapped to a score of two.

Table 9.1 Minimum Number of Correct Responses Required for High School Equivalency Certificate

Subtest	Number-Correct	Scaled Score
Reading	21	8
Writing	20	6
Mathematics	19	8
Science	20	8
Social Studies	20	8

Note that the corresponding scaled score for Writing is 6 rather than 8. This is because the High School Equivalency threshold for the essay score was set at 2. The sum of the two Writing components equaled the desired threshold value (6+2=8).

Another fixed point was then established on each score scale as the level at which test takers were judged as "college and/or career ready." Although this designation has a variety of definitions, it is generally understood to imply that a test taker has the mathematics and language skills necessary to qualify for and succeed in entry-level college courses without need of remediation. The college and career readiness threshold for each HiSET subtest was set at 15.

College and/or career readiness (CCR) performance levels and corresponding base-form number-correct scores were determined by taking advantage of the close relationships that HiSET shares with two other national testing programs. The first of these is the Iowa Tests of Educational Development (ITED), which was one of the sources of both content specifications and item content for the HiSET exam. The second is the ACT® college admissions test, on which college and/or career readiness thresholds have long been established (Allen & Sconing, 2005; Allen, 2013). Threshold scores (or *benchmarks*) have been set for each of the four tests that constitute the ACT battery — Mathematics, Science, Reading, and English. These scores were informed first by expert judgment through standard-setting procedures much like those that determined the HiSET passing thresholds. Judgments were supplemented by empirical longitudinal analyses that related high school ACT scores to college success (Allen, Radunzel & Moore, 2017). More specifically, students achieving the benchmark ACT scores have a 50% chance of receiving a grade of B or better and an 80% chance of earning a C or better in certain first-year courses. The benchmark scores for each ACT test are listed below:

Table 9.2 Benchmark Scores for the ACT

ACT Test	Benchmark Score
Reading	22
Writing	18
Mathematics	22
Science	23

The ITED is administered in conjunction with the ACT to large samples of Iowa high-school students, allowing links to be drawn between the score scales of these two programs. Although concordance links have been established between ACT and ITED tests by a number of researchers using a variety of methodologies (see Yin, Brennan, & Kolen, 2004), those produced by Furgol, Fina, & Welch (2011) are ideal for the current purposes as they are based on recent data and focused specifically on the correspondence of the ACT benchmark scores to their nearest ITED counterparts in terms of content.

The ACT – ITED concordances produced by Furgol, Fina, & Welch (2011) were based on samples of 14,000 to 18,000 Iowa high-school students who took both the ACT and ITED test batteries between 2007 and 2008. Although the content and item types of the ACT and ITED tests aligned with one another reasonably well, the correlations between their scores were only modest, ranging from a low of .68 for Science to .75 for Reading, Mathematics and English. However, these correlations were attenuated both because the two test batteries were administered at different points in the school year and due to restriction of range. While the ITED was administered to all Iowa high school students, only about half of those students also took the ACT, with this half being substantially more able and less variable than the whole. Adjusting for restriction of range increased the ACT – ITED correlations to .81–.83 for Reading, Mathematics, and English, and .76 for Science.

The Iowa data were used to translate ACT benchmarks to ITED score values by several methods, with the authors suggesting that the equal-error method produced the most satisfactory results. This method first classifies students on the basis of their ACT scores as above or below the benchmark threshold. It then computes for each ITED scale score value the *specificity* and *sensitivity* rates. The specificity rate is the proportion of students above the specified ITED score value who are also above the ACT benchmark.

Correspondingly, the sensitivity rate is the proportion of students below the ITED score value who also fall short of the ACT benchmark. Computing these rates across the range of ITED scores produces a pair of curves. The specificity curve begins near zero for low ITED score values (since students scoring at low ITED levels are unlikely to have ACT scores exceeding the benchmarks) and rises as ITED scores increase. The sensitivity curve starts near 1.00 for low ITED values (since students at those ITED levels likely have ACT scores that fall short of the benchmark threshold) and falls as ITED scores increase. The two curves cross at some ITED score, where specificity and sensitivity values are equal. The equal-error rate method chooses this value as the translation of the ACT benchmark.

Table 9.3 ACT Benchmark Scores and the Corresponding ITED Score

Subtest	ACT Benchmark Score	ITED Score
Reading	22	302
English	18	293
Mathematics	22	312
Science	23	329

The content and item formats of the HiSET Reading, Mathematics, and Science tests align closely with both their ACT and ITED counterparts. In fact, the HiSET base form items had initially been developed for use on ITED forms. The selected-response component of the HiSET Writing test also aligns closely with the ACT and ITED English tests. However, there is no exact match on the ACT battery to the HiSET Social Studies test. Neither has a link been established between the ITED Social Studies test and any ACT CCR benchmark. As such, the CCR threshold for HiSET Social Studies can be only roughly approximated, as described below.

The ACT/ITED CCR benchmarks can be directly extended to the HiSET score scale for the Reading, Writing, Science, and Mathematics content areas because each item on the HiSET base forms was administered alongside an ITED administration. Thus, a sample of test takers took one of the HiSET base forms and the ITED. This allows the items from both tests to be calibrated by item response theory methods on the same proficiency metric. Number-right scores on the HiSET base forms can then be mapped through the ITED proficiency scale to ITED scaled scores by the same methods, described below, that are used to link new HiSET forms to their respective base forms. Base form number-right scores corresponding to ITED/ACT CCR benchmarks are shown on the following page in Table 9.4.

Also of note is the HiSET Social Studies test, for which no corresponding ACT/ITED benchmark score exists. The HiSET CCR threshold was therefore set by observing that the thresholds established for the other four content areas stood at somewhere between the 70th and 75th percentile of the first-year HiSET score distribution (based on those candidates who tested in the first half of 2014). The Social Studies CCR threshold was accordingly set at the 75th percentile as well, resulting in a base form number right score of 34.

Table 9.4 Minimum Number of Correct Responses on Base Forms Required for College and/or Career Readiness Designation				
Content Area	ACT Benchmark Score	ITED Score	HiSET Base Form Number Right	HiSET Scaled Score
Reading	22	302	35	15
English / Writing	18	293	38	11
Mathematics	22	312	33	15
Science	23	329	34	15
Social Studies	N/A	N/A	34	15

Note that the corresponding scaled score for the HiSET Writing subtest is 11 rather than 15. This is because the college-readiness threshold for the essay score was set at 4. The sum of the two Writing components again equaled the desired threshold value (11+4=15).

Setting average scores equal across all HiSET subtest scaled scores introduces a third fixed point on each scale. Average scores were first computed across nearly 13,000 test takers who tested under operational conditions during the first six months of 2014, a period that preceded the establishment of the current score scale. The average score on each base form (except Writing) was mapped to a scaled score of 11. The Writing average base form score was mapped to a scaled score of 8 because the average essay score was approximately three, meaning that the total Writing scaled score would be 11, matching the other subtests.

Table 9.5 Average Number-Correct Scores of Test Takers Tested in Early 2014			
Subtests	Base Form	Average Number-Correct	Scaled Score
Reading	A	26.4	11
Writing	A	29.9	8
Mathematics	A	25.5	11
Science	A	26.3	11
Social Studies	C	26.0	11

After determination of the three scale points described above (passing threshold, college-readiness threshold and average score), the remaining number-right scores on each base form were mapped to their associated scaled score values. This was done with regard to two additional principles. The first was that scores should be maximally distinguishable in the neighborhood of the passing threshold. In practice, this meant that base number-correct scores just below and just above the threshold mapped to scaled scores of 7 and 9, respectively. This left only a single number-right score mapping to the passing threshold whereas other scaled-score values were mapped to by multiple number-right scores.

The second principle held that scaled-score distributions should be as similar in shape as possible across subtests. Although the average scores were already fixed as equal, it is also desirable that standard deviations and higher moments of the distributions be similar as well. Meeting both of the goals would mean that a given score would have similar percentile ranks regardless of the subtest in which it was achieved.

9.3 Maintaining the HiSET Score Scale across Test Forms

Test equating methods have been developed to statistically adjust scores to account for the fact that no two test forms are exactly equal in difficulty (Kolen & Brennan, 2014). The purpose of these methods is to ensure that scaled scores are comparable across test forms even when number-correct scores are not. Because multiple HiSET forms are produced and administered interchangeably, it is important that a test taker's score not depend strongly on the particular form he or she was administered. The scaling methods described above determined how number-correct scores on the base test form mapped to scaled scores. Equating methods, in contrast, determine how number-correct scores on new test forms convert to equivalent number-correct scores on the base form and, by extension, to scaled scores.

The necessity for equating can be illustrated by an example. Consider two forms of the HiSET Mathematics test, each containing 50 items and each measuring the same concepts in very similar ways such that teachers or other subject-matter experts reviewing the items conclude that the forms are substantively equivalent. Although substantive equivalence is a necessary condition for scores to be comparable across test forms it is not sufficient. Suppose that one of the two Mathematics forms contained items that were, on average, slightly more difficult than those on the other form. This means that number-correct scores on the two forms are not directly comparable. For example, if the two test forms were administered to two groups of test takers that were equivalent in background, educational achievement, level of preparation and all other important ways, the resulting distributions of number-correct scores would differ, with the average score for the easier form being higher than the average on the more difficult form.

Test equating methods allow the number-correct scores on different forms to be statistically adjusted in ways that make them more comparable. The result is a table that maps each score on a "new" test form to the most comparable score on the base form. For example, in the case above, a score of 29 on the easier Mathematics form might equate to a score of 27 on the harder form. Suppose further that the harder form was designated as the base form and that a number-correct score of 27 was associated with a scaled score of 12. Then, by extension, a score of 29 on the new and easier form would also be associated with a scaled score of 12. Completing this table across all number-correct scores on the new form allows scaled scores achieved on that form to be equivalent to scaled scores achieved on the base form. Although the number-correct score distributions differed between the easier and harder forms in the example above, equating would result in scaled score distributions that are much more similar.

A full description of the equating methods used to ensure that all HiSET forms produce comparable scaled scores requires the following elements:

(1) A description of the equating methods.

(2) A description of the data to which these methods are applied.

(3) A description of the operational procedures employed.

Each of these elements is described on the following pages.

9.3.1 Equating Methods

New HiSET test forms are equated to base forms through methods known as *item response theory true score equating*. The details of these methods are beyond the scope of this technical report, but see Kolen & Brennan, 2014 for a complete description.

Item response theory (IRT) equating methods offer a compelling advantage over other methods in allowing newly-developed test forms to be *pre-equated*. This means that the tables that convert number-correct scores on a new form to base-form equivalents (and to scaled scores) can be computed before the new form has been administered, thus allowing scaled scores to be produced and reported on the new form immediately upon its release for operational use.

The new HiSET forms introduced each year are pre-equated so that all candidates can have their scores reported in a timely manner. Other forms of equating require that substantial data samples be collected on the new form before the score conversion tables can be determined. Use of these methods would have therefore required that scoring of newly-introduced forms be delayed until data were collected and conversion tables produced.

The key requirement that allows pre-equating is that IRT item parameter estimates be known for all items on the new form at the time it is assembled. How these item parameters estimates are obtained is described on page 64 in Section 9.3.2.

How IRT true score equating allows score conversion tables to be produced prior to a newly-developed test form being administered is also best illustrated by a diagram and an example. Each item on the new test form has an associated item response function, as determined by the item parameters estimated uniquely for that item. Item responses functions can be summed across the items on the test form, producing a *test characteristic curve* (TCC), as shown in Figure 9.1.

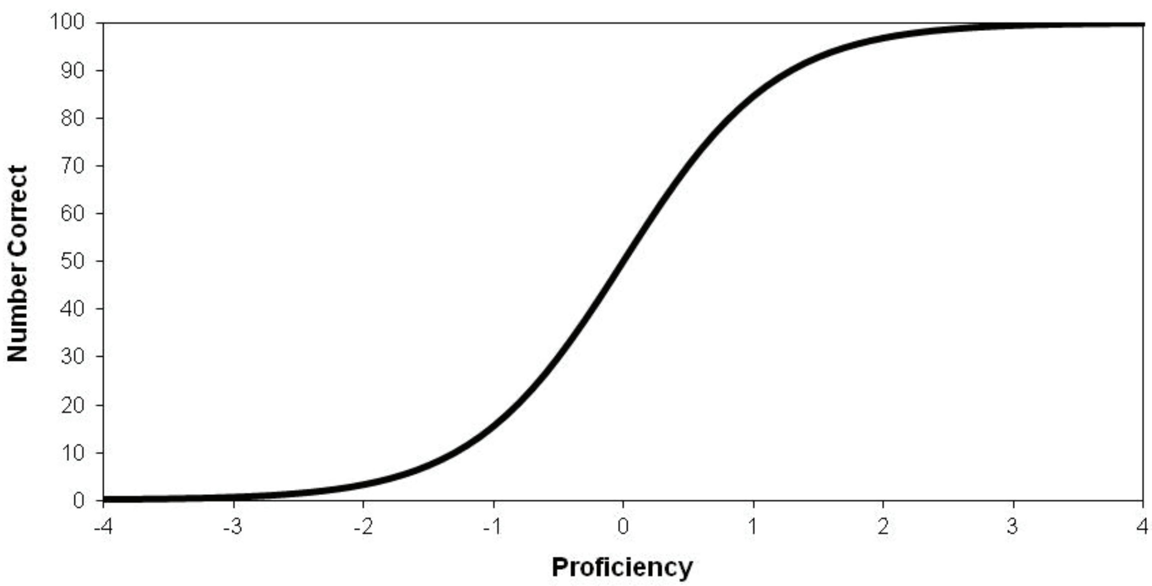

Figure 9.1 Test characteristic curve.

While the item response function relates test taker proficiency (or theta) with the probability of a correct response, the TCC relates proficiency to a test taker's expected number-correct score. This is because an expected number-correct score is simply the sum of the probabilities of answering each individual item correctly.

Suppose that there are now TCCs estimated for both the base test form and a newly-assembled form that happens to be much more difficult than the base form. These TCCs are both depicted in Figure 9.2. The key assumption made in plotting both the new form and base TCC on the same proficiency scale is that the parameter estimates for all items on both tests share the same scale. The procedures for ensuring that this is the case will be described below.

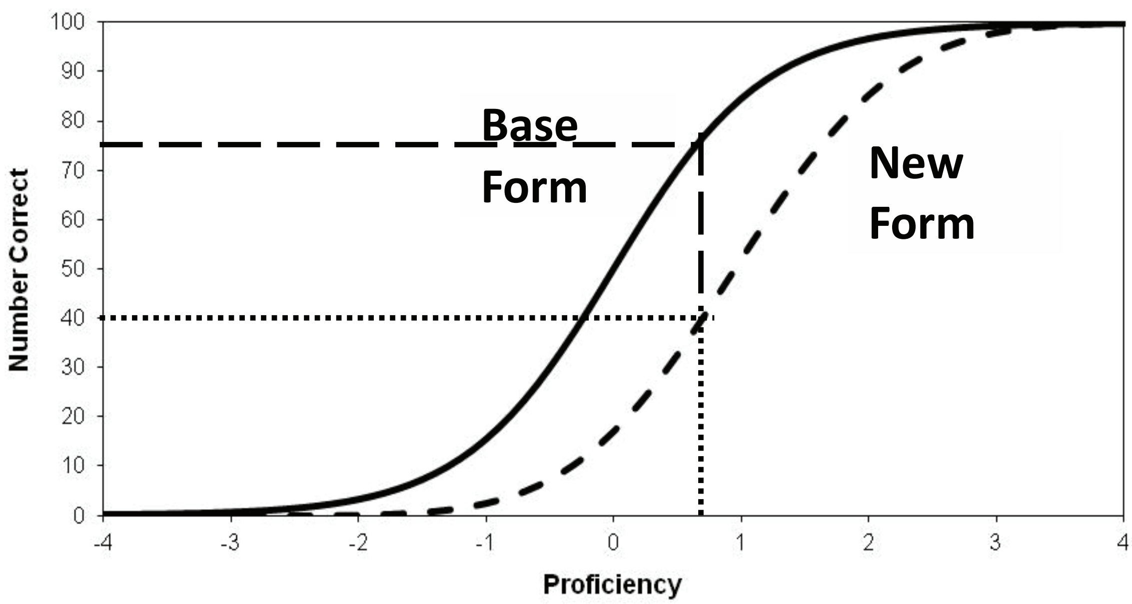

Figure 9.2 Base and new form TCCs.

Consider a score of 40 on the new form. The proficiency value associated with this score is determined by finding the value on the horizontal (proficiency) axis immediately beneath the point where the new form TCC crosses 40 on the vertical (score) axis. This value is 0.7. Then find the point on the base form TCC associated with a proficiency of 0.7 and follow it to the vertical axis to find the corresponding number-correct score (75). Since both scores of 40 on the new form and 75 on the base form are the expected results of test takers with proficiency equal to 0.7, they represent equivalent levels of performance. Repeating this process for all possible number-correct scores on the new form completes the conversion table that translates new form scores to base form equivalents. These base form equivalent scores can then be transformed to scaled scores by applying the raw-to-scale conversion for the base form. It should be noted that this example was exaggerated to show how the scores on two forms are equated. The differences in the scores across forms on the HiSET subtests is typically closer than in this example.

9.3.2 Pretest Data Collection

As noted, the key requirement of the IRT true score equating method is that each item on the new form has estimated item parameters that are comparable to those from the items on the base form. These parameters are estimated from pretest data collected prior to assembly of the new form. Pretest data are currently collected through the HiSET program's close association with the Iowa Tests of Educational Development (ITED) (Feldt, Forsyth, & Alnot, 1986). The ITED is administered each year to large samples of high school juniors and seniors, a population appropriate for estimating performance of HiSET items. Each year, newly-developed HiSET items are embedded within the ITED and administered to samples of 2,000–3,000 test takers. Although the HiSET items do not contribute to ITED scores, their location within the ITED forms is not made known, ensuring that test takers are motivated to perform to the best of their abilities. The data collected are used first to evaluate item quality, with poorly performing items eliminated from future use on operational HiSET forms. The data from the surviving items are calibrated to obtain their item parameter estimates.

9.3.3 Item Calibration and Scale Linking

HiSET items are calibrated under the three-parameter logistic model (3PL) along with the ITED items that comprise the operational forms of the ITED. The software routine BILOG MG-3 (Zimowski et al., 2003) is used to conduct the calibrations, with default prior distributions imposed on the *a* and *c* parameters (Zimowski et al., 2003, p. 187). The resulting estimates are inspected to ensure appropriate levels of model-data fit (Ames & Penfield, 2015). A statistical approach was used to evaluate model-data fit. The chi-square values were evaluated by calculating the adjusted fit values and flagging items with adjusted fit values greater than 0.45, following the classification by Cohen (Cohen, 1988, pp. 224–225). The adjusted fit values were calculated by dividing the chi-square fit statistic by the sample size using the following formula (Barton & Huynh, 2003):

$$C = \sqrt{\frac{X^2}{X^2 + N}} \tag{9-1}$$

Appendix C, Tables C.1 to C.5 present summaries of the parameter estimates, by form, for each HiSET subtest.

As estimated, pretest item parameters are not necessarily on the same proficiency scale as the HiSET base form parameters. It is therefore necessary to *rescale* or *link* the pretest item parameter estimates to place them on the base form scale. To do so requires a set of *anchor items* which have previously been linked to the base form scale.

Anchor items serve as bridge between the pretest and base scales. Anchors items are administered and calibrated along with the pretest items, with their newly-estimated item parameters taking their place on the pretest scale. However, anchor items also have another set of estimated values from a previous administrations, and these estimates have already been linked to the base scale (by the same methods and process described below). For the HiSET program's purposes, a designated set of operational ITED items serve as anchors.

Scale linking is the process of a finding a linear transformation that adjusts the new (pretest) parameter estimates of the anchor items to be maximally similar (in some sense) to the old (base) parameter estimates of these same items. Different scale linking methods use different definitions of "maximally similar." The

HiSET program uses the Stocking-Lord method that minimizes the weighted squared difference between the TCCs computed from the new and old parameter estimates for the anchor items (Stocking & Lord, 1983). A software package called STUIRT (Hanson, Zeng, & Cui, 2004) is used to estimate the linking relationship and adjust the pretest items to the base form scale. Once on the base form scale, parameter estimates for the newly-developed items can indeed be properly used in the equating procedures illustrated in Figure 9.2.

Figures 9.3 through 9.7 present TCCs for each subtest. The curves in each figure are for the three new or pretest forms (Forms A, B, and C) administered in 2015 and the old or base form administered in 2014.

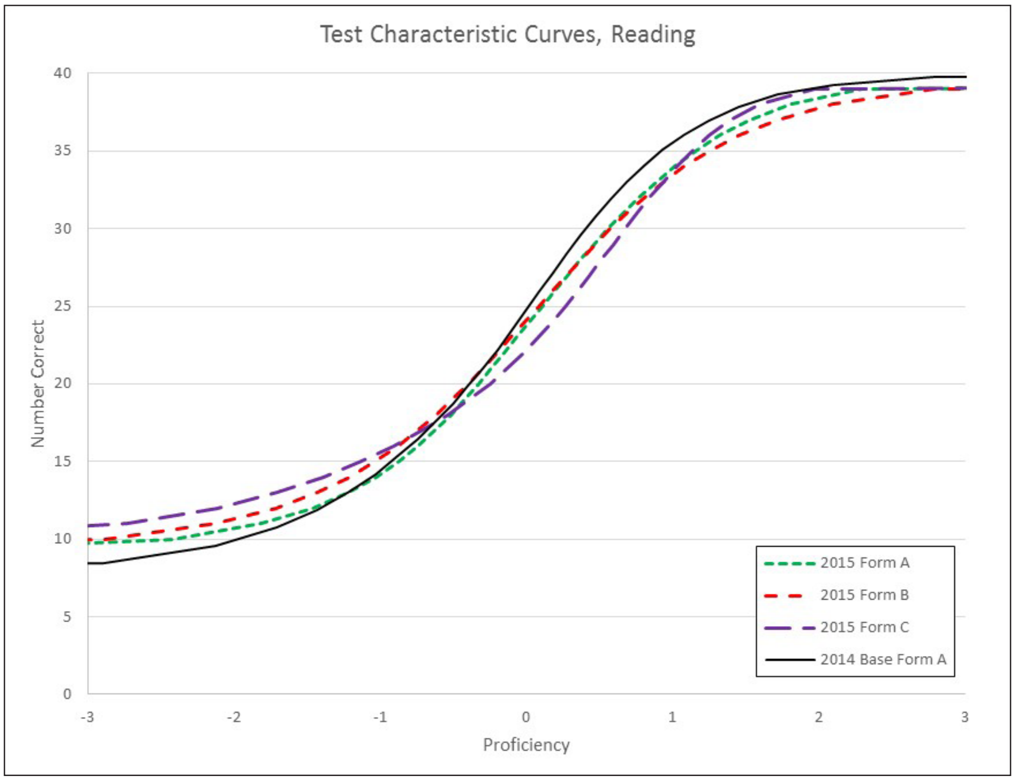

Figure 9.3 Test characteristic curves for Reading.

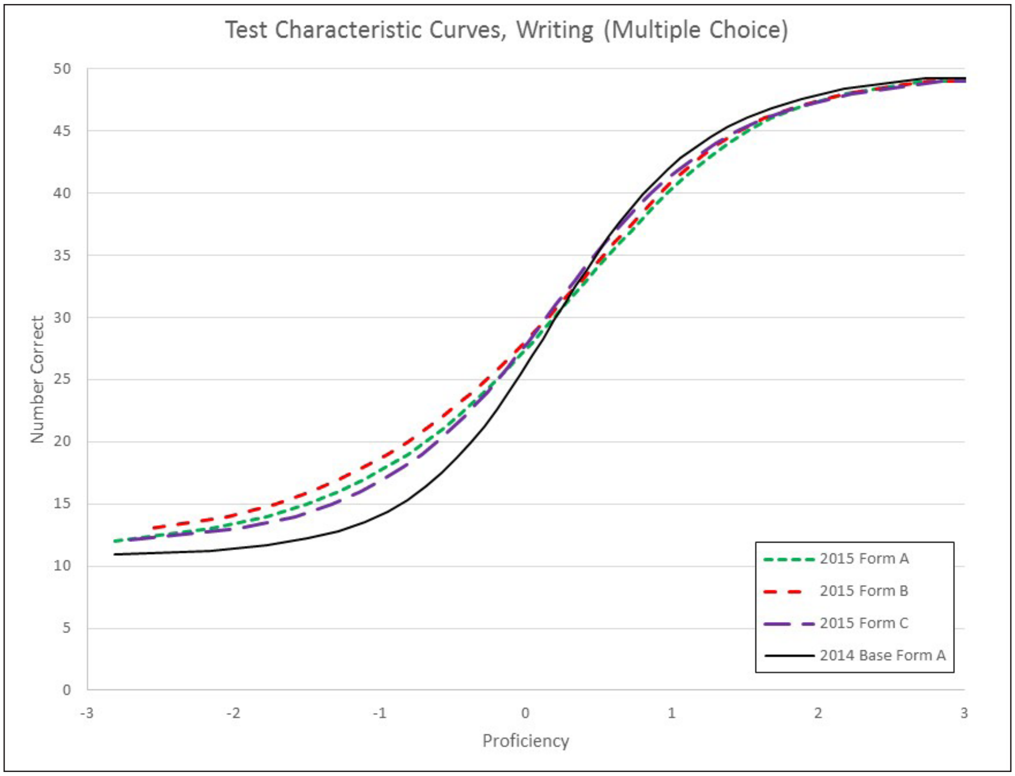

Figure 9.4 Test characteristic curves for Writing (multiple-choice items only).

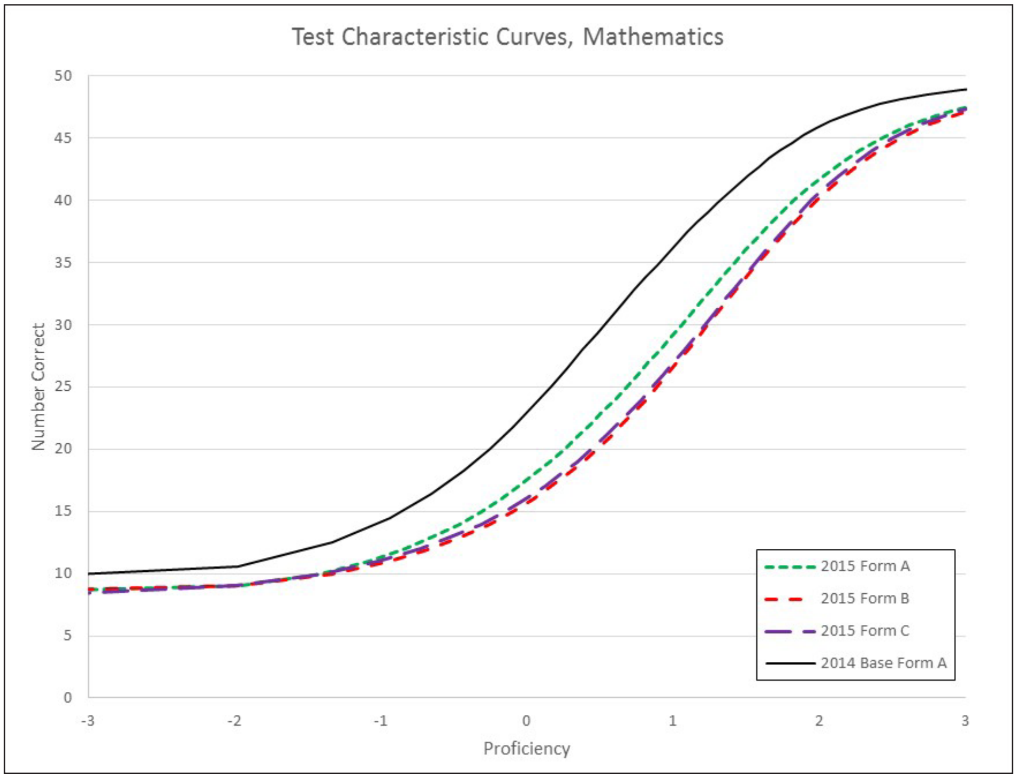

Figure 9.5 Test characteristic curves for Mathematics.

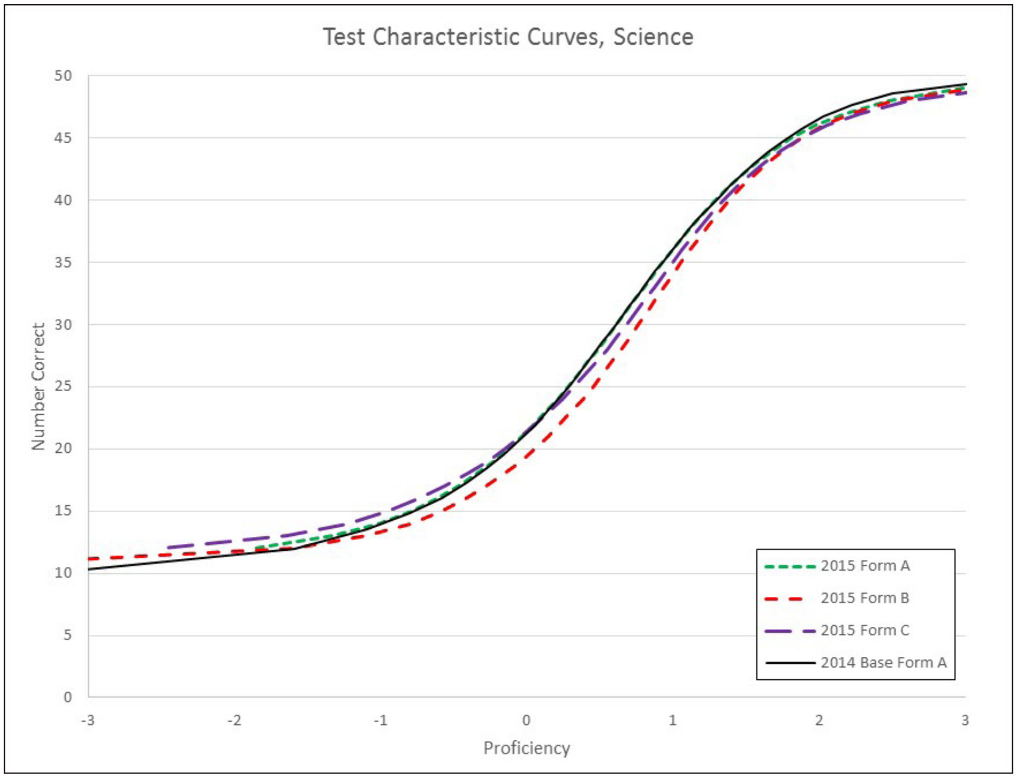

Figure 9.6 Test characteristic curves for Science.

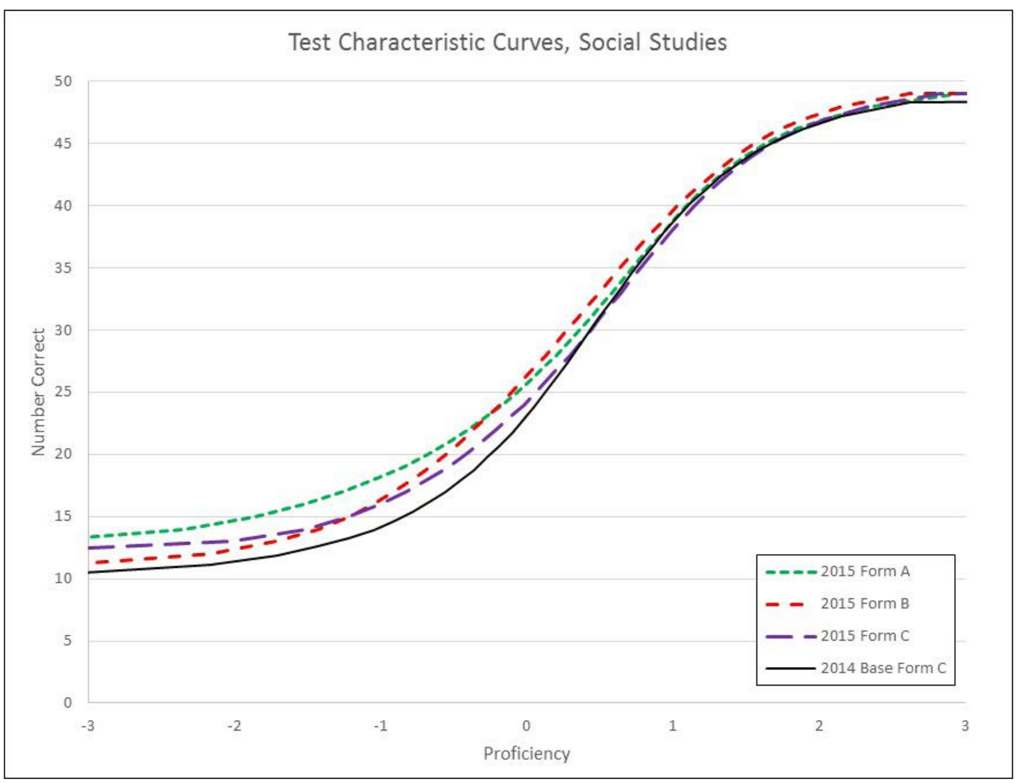

Figure 9.7 Test characteristic curves for Social Studies.

Note that, unlike the TCCs for the other subtests, the TCCs for the new Mathematics forms do not closely align with the base form (see Figure 9.5). This is due to the fact that the content specifications for Mathematics were adjusted in 2015 to include a few more Algebra items which were relatively difficult. Following the 2015 administration, test developers had access to a larger pool of Algebra items; consequently, it is expected that future Mathematics forms will not be as difficult.

9.4 Quality Control Procedures

To ensure that the above procedures are properly applied, a series of quality control checks are routinely conducted. These checks make certain both that the pretest data samples are properly coded and structured and verify that analyses of these data produce appropriate results. These checks include:

- Verification that the contents, coding and layout of the pretest data files meet specifications.
- Confirming that the scoring of all items and tests is correct.
- Confirming that the item parameter estimates fall within expected ranges and that model-data fit is acceptably high.
- Confirming that scale linking transformation values fall within expected ranges.
- Confirming that the number-correct score to scaled score conversion tables for each new test form fall within expected ranges.

Chapter 10: Test Taker Performance

10.1 Scale Score Results

Test takers' total test scores are scale scores derived from the IRT procedures described in Chapter 9. Overall summary statistics and performance level results provided in Chapter 10 are for **English online test takers only**. Performance level results are presented by gender and race/ethnicity. Similar information based on the English paper, Spanish online, and Spanish paper test taker results are presented in Appendix D, specifically Tables D.1 to D.10 for **English paper-based test takers**, Tables D.11 to D.20 for **Spanish online test takers**, and Tables D.21 to D.30 for **Spanish paper-based test takers**.

Table 10.1 provides scale score summary statistics based on all test takers who took the online English-language version of the 2015 subtests. The information is presented for all three forms of a subtest combined and for each form separately. The observed mean scale scores ranged from 8.7 for Mathematics Form B to 13.39 for Science Form C. Although the HiSET scales were developed in 2014 so that the mean scale scores were the same across the subtests, the observed mean scale scores reflect the difficulty of the items reported in Table 5.1. As observed in Table 5.1, the half of the Mathematics items have a p-value less than .40. The mean p-values and the lower mean scale scores indicate that the Mathematics subtest was challenging for the test takers. Scale score summary statistics for the Writing CR prompts, by form, are presented in Table 10.2. The mean CR scale scores ranged from 3.11 for Form A Prompt 1 to 3.32 for Form B Prompt 3; the median values ranged from 3 to 3.5 across all CR items. The distributions of essay scale scores, by form and by prompt, are presented in Tables 10.3 to 10.5. Across all 8 writing prompts, the majority of test takers received a scale score of 3 or 4.

Table 10.1 Total Test Scale Score Summary Statistics, Overall and by Form: English, Online Test Takers

		N	Mean	*SD*	Median
Reading	Overall	26,574	12.12	3.99	13
	Form				
	A	9,045	12.21	3.85	13
	B	8,910	12.26	3.97	13
	C	8,619	11.86	4.12	12
Writing	Overall	24,223	12.22	3.12	12
	Form				
	A	6,142	12.39	3.07	13
	B	9,076	12.24	3.17	12
	C	9,005	12.09	3.11	12
Mathematics	Overall	30,402	9.17	3.89	9
	Form				
	A	10,149	9.68	3.61	9
	B	9,937	8.70	4.21	8
	C	10,316	9.11	3.75	9
Science	Overall	24,998	13.02	3.87	13
	Form				
	A	8,349	12.61	3.71	13
	B	8,414	13.07	3.79	14
	C	8,235	13.39	4.08	14
Social Studies	Overall	26,887	11.71	4.24	11
	Form				
	A	8,980	11.73	4.28	11
	B	8,962	11.84	4.2	11
	C	8,945	11.57	4.24	12

Table 10.2 Scale Score Summary Statistics for Writing CR Prompts, by Form: English, Online Test Takers

CR Prompt	*N*	Mean	Median	Standard Deviation
Form A, Prompt 1	3,053	3.11	3.0	0.89
Form A, Prompt 2	3,020	3.19	3.0	0.87
Form B, Prompt 1	2,952	3.29	3.5	0.87
Form B, Prompt 2	3,007	3.17	3.0	0.87
Form B, Prompt 3	3,044	3.32	3.5	0.88
Form C, Prompt 1	2,956	3.26	3.0	0.91
Form C, Prompt 2	2,990	3.25	3.0	0.86
Form C, Prompt 3	3,004	3.29	3.5	0.82

Table 10.3 Scale Score Distributions by Prompt, Writing Form A: English, Online Test Takers

Essay Score	Prompt 1		Prompt 2	
	Number of Test Takers	Percent of Test Takers	Number of Test Takers	Percent of Test Takers
0	39	1.3	30	1.0
1	152	4.9	109	3.6
2	389	12.6	325	10.7
3	1,207	39.0	1,230	40.3
4	1,075	34.8	1,105	36.2
5	206	6.7	210	6.9
6	24	0.8	41	1.3

Table 10.4 Scale Score Distributions by Prompt, Writing Form B: English, Online Test Takers

Essay Score	Prompt 1		Prompt 2		Prompt 3	
	Number of Test Takers	Percent of Test Takers	Number of Test Takers	Percent of Test Takers	Number of Test Takers	Percent of Test Takers
0	26	0.9	27	0.9	20	0.7
1	92	3.1	108	3.6	77	2.5
2	267	9.0	347	11.4	308	10.1
3	1,099	36.9	1,227	40.4	1,059	34.6
4	1,167	39.2	1,059	34.9	1,249	40.8
5	289	9.7	232	7.6	299	9.8
6	38	1.3	34	1.1	52	1.7

Table 10.5 Scale Score Distributions by Prompt, Writing Form C: English, Online Test Takers						
	Prompt 1		Prompt 2		Prompt 3	
Essay Score	Number of Test Takers	Percent of Test Takers	Number of Test Takers	Percent of Test Takers	Number of Test Takers	Percent of Test Takers
0	18	0.6	23	0.8	14	0.5
1	109	3.7	84	2.8	65	2.2
2	317	10.7	312	10.4	285	9.4
3	1,061	35.7	1,159	38.5	1,146	38.0
4	1,148	38.6	1,134	37.6	1,237	41.0
5	269	9.0	267	8.9	250	8.3
6	52	1.7	34	1.1	21	0.7

Total test scale score analyses were also conducted to investigate test taker performance by groups of interest. Tables 10.6 to 10.10 provide summary statistics by gender and by race/ethnicity groups. On average, males performed slightly higher across the five subtests compared to females, with the exception of Writing in which females scored slightly higher. Some variability in performance was observed across the race/ethnicity groups for all subtests.

Table 10.6 Total Test Scale Score Summary Statistics for Reading, by Demographic Group: English, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Gender	Male	13,992	53	12.23	13	4.02	1	20
	Female	12,582	47	11.99	12	3.94	1	20
Race/ Ethnicity	American Indian	330	1	11.45	12	4.23	2	20
	Asian	455	2	9.63	9	4.65	1	20
	African American	4,614	17	9.91	10	3.74	1	20
	White	13,209	50	13.28	14	3.64	1	20
	Hispanic	4,282	16	11.18	11	3.78	1	20
	Pacific Islander	49	> 1	11.67	13	4.36	2	19
	Multiracial	903	3	13.04	13	3.55	2	20
	Other/ No Response	2,732	10	11.90	12	4.08	1	20

Table 10.7 Total Test Scale Score Summary Statistics for Writing, by Demographic Group: English, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Gender	Male	12,978	54	11.98	12	3.17	1	20
	Female	11,245	46	12.50	12	3.04	1	20
Race/ Ethnicity	American Indian	295	1	11.28	11	3.27	3	19
	Asian	365	2	11.45	11	3.79	2	20
	African American	3,876	16	10.84	11	2.92	1	20
	White	12,658	52	12.84	13	3.05	1	20
	Hispanic	3,668	15	11.63	12	2.88	2	20
	Pacific Islander	44	> 1	12.14	12	2.91	5	18
	Multiracial	870	4	12.97	13	3.03	2	20
	Other/ No Response	2,447	10	12.04	12	3.13	2	20

Table 10.8 Total Test Scale Score Summary Statistics for Mathematics, by Demographic Group: English, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Gender	Male	15,482	51	9.73	10	3.93	1	20
	Female	14,920	49	8.59	8	3.75	1	20
Race/ Ethnicity	American Indian	373	1	8.61	8	3.77	1	19
	Asian	455	2	10.27	10	5.00	1	20
	African American	5,737	19	7.34	7	3.34	1	20
	White	14,820	49	10.00	10	3.87	1	20
	Hispanic	4,798	16	8.59	8	3.55	1	20
	Pacific Islander	66	> 1	8.36	9	3.46	1	15
	Multiracial	1,052	3	9.98	10	3.92	1	20
	Other/ No Response	3,101	10	9.12	9	3.88	1	20

Table 10.9 Total Test Scale Score Summary Statistics for Science, by Demographic Group: English, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Gender	Male	13,127	53	13.60	14	3.84	1	20
	Female	11,871	47	12.38	13	3.81	1	20
Race/ Ethnicity	American Indian	302	1	12.45	13	3.96	2	20
	Asian	371	1	12.30	12	4.26	2	20
	African American	4,277	17	10.52	10	3.52	1	20
	White	12,700	51	14.20	15	3.53	1	20
	Hispanic	3,880	16	11.87	12	3.66	1	20
	Pacific Islander	46	> 1	12.30	13	3.92	3	18
	Multiracial	876	4	14.05	14	3.39	2	20
	Other/ No Response	2,546	10	12.91	13	3.91	1	20

Table 10.10 Total Test Scale Score Summary Statistics for Social Studies, by Demographic Group: English, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Gender	Male	13,867	52	12.46	13	4.25	1	20
	Female	13,020	48	10.92	10	4.09	1	20
Race/ Ethnicity	American Indian	340	1	11.09	11	4.26	2	20
	Asian	417	2	10.77	10	4.51	2	20
	African American	4,752	18	9.31	9	3.61	1	20
	White	13,366	50	12.91	13	4.09	1	20
	Hispanic	4,354	16	10.56	10	3.90	1	20
	Pacific Islander	47	> 1	11.55	11	3.70	4	19
	Multiracial	909	3	12.86	13	3.96	3	20
	Other/ No Response	2,702	10	11.74	11	4.25	2	20

10.2 Performance Level Results

As described in the introduction of this technical report, the results of the HiSET subtests are used to give out-of-school youth and adults the best opportunity to demonstrate their skills and knowledge and earn a state-issued high school equivalency credential.

Performance on each of the five HiSET subtests results in a scale score between 1 and 20. A score of at least 8 on each MC test and 2 on the essay component of the Writing test is required to pass the HiSET test and be certified as performing at a level consistent with high school completion equivalency. The HiSET test also has a cut score to indicate that the test taker has performed at the College and Career Readiness (CCR) level. A scale score of at least 15 on each of the five multiple-choice tests and at least 4 on the essay component of the Writing test are required to demonstrate CCR. Tables 10.11 through 10.15 present the percentages of test takers identified in each of three performance levels:[4]

- Did not pass high school equivalency ("Did Not Pass"),
  - Test taker demonstrates minimal understanding of the subject and has not demonstrated the ability to apply the knowledge and skills that are associated with high school graduation requirements.

[4] Performance level categories were defined during standard setting in 2014.

- Passed high school equivalency ("Passed But Not CCR"),
  - Pass but not College and Career Ready — Test taker demonstrates adequate understanding of the subject and has the ability to apply the knowledge and skills that are associated with high school graduation requirements.
- Passed college and career readiness ("College & Career Ready"),
  - College and Career Ready — Test taker demonstrates thorough understanding of the subject and has the ability to apply the knowledge and skills that are associated with readiness for college and various career paths.

This information is presented within each subtest for total test takers, and by gender and race/ethnicity. As shown in Tables 10.11 through 10.15, the percentages of test takers who passed the HiSET subtest, but were not CCR, ranged from 53.38% for Science to 70.18% for Writing. The percentages of test takers who were CCR ranged from 9.36% for Mathematics to 28.19% for Science.

Table 10.11 Percentage of English, Online Test Takers in each Performance Level: Reading

	N	Percent of Total	Performance Levels		
			Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	26,574		14	59	27
Gender					
Male	13,992	53	13	58	28
Female	12,582	47	14	60	26
Race/Ethnicity					
American Indian	330	1	20	59	21
Asian	455	2	36	49	16
African American	4,614	17	27	63	10
White	13,209	50	7	56	37
Hispanic	4,282	16	17	65	17
Pacific Islander	49	> 1	18	55	27
Multiracial	903	3	7	60	33
Other	2,732	10	16	59	26

Note. Test takers who chose not to select one of the specific responses to the Race/Ethnicity questions are classified as 'Other.'

Table 10.12 Percentage of English, Online Test Takers in each Performance Level: Writing					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	24,223		9	70	21
Gender					
Male	12,978	54	11	70	19
Female	11,245	46	6	70	24
Race/Ethnicity					
American Indian	295	1	16	69	15
Asian	365	2	19	61	20
African American	3,876	16	15	76	9
White	12,658	52	6	67	27
Hispanic	3,668	15	10	76	15
Pacific Islander	44	> 1	11	70	18
Multiracial	870	4	6	66	28
Other	2,447	10	9	71	19

Note. Test takers who chose not to select one of the specific responses to the Race/Ethnicity questions are classified as 'Other.'

Table 10.13 Percentage of English, Online Test Takers in each Performance Level: Mathematics					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	30,402		37	54	9
Gender					
Male	15,482	51	31	57	12
Female	14,920	49	42	51	7
Race/Ethnicity					
American Indian	373	1	42	50	7
Asian	455	2	36	42	22
African American	5,737	19	56	41	3
White	14,820	49	28	59	13
Hispanic	4,798	16	41	54	6
Pacific Islander	66	> 1	44	52	5
Multiracial	1,052	3	30	57	13
Other	3,101	10	38	53	9

Note. Test takers who chose not to select one of the specific responses to the Race/Ethnicity questions are classified as 'Other.'

Table 10.14 Percentage of English, Online Test Takers in each Performance Level: Science					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	24,998		9	53	38
Gender					
Male	13,127	53	7	49	44
Female	11,871	47	10	59	31
Race/Ethnicity					
American Indian	302	1	13	55	32
Asian	371	1	14	53	34
African American	4,277	17	19	67	13
White	12,700	51	4	46	50
Hispanic	3,880	16	11	64	25
Pacific Islander	46	> 1	11	61	28
Multiracial	876	4	3	51	45
Other	2,546	10	10	54	37

Note. Test takers who chose not to select one of the specific responses to the Race/Ethnicity questions are classified as 'Other.'

Table 10.15 Percentage of English, Online Test Takers in each Performance Level: Social Studies					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	26,887		18	53	28
Gender					
Male	13,867	52	14	51	35
Female	13,020	48	23	56	21
Race/Ethnicity					
American Indian	340	1	24	52	24
Asian	417	2	25	53	22
African American	4,752	18	35	56	10
White	13,366	50	11	51	38
Hispanic	4,354	16	24	58	18
Pacific Islander	47	> 1	19	53	28
Multiracial	909	3	10	53	37
Other	2,702	10	18	54	28

Note. Test takers who chose not to select one of the specific responses to the Race/Ethnicity questions are classified as 'Other.'

Chapter 11: Quality Control Procedures

11.1 Quality Control of Test Materials

ETS follows a set of internal quality standards to ensure high-quality online published products for all testing-related materials. Quality control in test administration requires that the contents of all test materials (including electronic information, ad hoc documents, and test administration manuals) align with one another and present accurate information because contradicting information creates frustration to the test users and may impact the validity of test score interpretation.

To help ensure consistency in test materials used for the HiSET program, the manuals (i.e., test administration manuals, training materials, and technical manual) and the digital information are reviewed by subject matter experts at ETS. Documents are developed through multiple iterations such as content review cycles and then undergo an editorial review by ETS internal editors. Reviews of the test materials are built into the planned test administration schedule and test materials are not released to the testing centers or the test users until after the HiSET program's final approval.

11.2 Quality Control of System Functionality

For the HiSET program, the ETS quality assurance team conducted testing procedures on the following aspects of the end-to-end system in both the user acceptance and production environments: test delivery and item content rendering. These activities adhere to the software development life cycle process as follows:

- Software Testing. ETS developed user acceptance test plans and test scripts. A number of testing activities took place with these plans and scripts, including the testing of software components, security testing, integration testing, hardware and network capacity testing, data conversion testing, and load testing.
- Data Conversion Testing. ETS performed testing for data conversions in the system. These data conversions include but are not limited to: test taker data import, test taker scoring, raw score test taker assignment, and raw score to individual test taker report conversion. Quality assurance professionals compared samples of data in order to verify that the source data matches the converted data in the destination systems.
- Hardware and Network Capacity Testing. ETS provided readiness tools to help districts and schools prepare their hardware and networks for the testing windows.
- System and Integration Testing. The ETS software quality assurance staff performed system-level testing. The staff validated the system against all requirements. This process included verifying system accessibility, links, scoring, reporting, security, and performance. During this phase, staff could detect and correct issues before the final release.
- Operational Trial. Prior to the release of each product, the ETS software quality assurance staff performed full system-level tests in an independent test environment that mirrors the production configuration. Staff members also tested the system on all supported computer platforms and browsers. These system-level tests included comprehensive assessments on functionality, usability, reliability, security, and overall performance. The staff verified that each webpage, link, item, and image displayed properly through the graphical user interface standards. During this process, the staff members also validated system content for accuracy.

- Load Testing. ETS regularly performed extensive load testing to determine system capacity and to provide quality delivery of online assessments. Load testing consisted of employing machines across the Internet to simulate the test taker testing environment. During this testing period, ETS staff obtained data that enables long-term scalability planning.
- Security Testing. In order to establish the integrity, confidentiality, and availability of the data, ETS used industry-standard tools to regularly run automated security scans against production and development networks and systems. Real-time vulnerability updates protected ETS systems against the very latest known threats.
- User Acceptance Testing. ETS developed and reviewed the user acceptance tests to confirm the system meets the requirements of the contract. In addition to the quality assurance checks on functionality, the system's consistency in capturing responses and transferring of test taker data for scoring was also evaluated.

For each of these quality checks, ETS staff members were required to evaluate specific functioning outlined on a quality assurance checklist.

11.3 Quality Control of Psychometric Analyses

ETS took various necessary measures to ascertain that the scoring keys were applied to the test taker responses as expected and the test taker scores were computed accurately. The psychometric analyses conducted at ETS underwent comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists were consulted by lead psychometricians to systematically review the statistical procedures performed on each HiSET subtest. Equatings and conversion tables were reviewed by two psychometricians before pre-equated test scores were released.

11.4 Quality Control of Scoring and Reporting

ETS's scoring and reporting systems have quality control procedures integrated throughout, including both automated and manual inspections, to ensure data accuracy. ETS Assessment Development, Research, and Statistical Analysis, Performance Assessment Scoring Service, and Information Technology groups all participated in certifying the scoring and reporting system to ensure operational readiness and scoring integrity. All teams followed established procedures required by the International Organization for Standardization (ISO) 9000 family of standards. The combined efforts of each of these groups provided multiple layers of quality assurance and control.

ETS built and reviewed the scoring system models based on the HiSET scoring specifications and requirements. Machine-scored item responses and demographic information were sent into a master test taker file. Test takers' essays were also sent electronically to the ETS Online Network for Evaluation (ONE) scoring centers for scoring by trained, qualified raters. Record counts were verified against the counts obtained during security check-in from the document processing staff to ensure all test takers were accounted for in the file.

Once the record counts were reviewed, the machine-scored item responses were scored against the appropriate approved answer key provided by the HiSET team. In addition, the test taker's original response string was stored for data verification and auditing purposes. ETS determined and refined the documentation of specifications for the scoring of answer documents well in advance of the receipt of test materials. These specifications contained detailed scoring procedures, along with the procedures for determining whether a test taker has attempted a test and whether that test taker should be included in statistics and calculations for computing summary data. Standard quality inspections were performed on all data files, including the evaluation of each test taker data record for correctness and completeness. Test taker results were kept confidential and secure at all times.

Upon the completion of the thorough data verification process, quality checks were performed on the data placement and report file formatting for each data element displayed on the reports. All reporting data elements were verified by comparing back to the production data file and the reporting processing rules. Additional quality crosschecks were performed to ensure accuracy and consistency across all reporting media for the assessment.

References

Allen, J. (2013). *Updating the ACT College Readiness Benchmarks*. Iowa City, IA: ACT.

Allen, J., Radunzel, J., & Moore, J. (2017). *Evidence for standard setting: Probabilities of success in "benchmark" college courses, by ACT test scores*. ACT Technical Brief. Iowa City, IA: ACT.

Allen, J., & Sconing, J. (2005) *Using ACT Assessment scores to set benchmarks for college readiness*. Iowa City, IA: ACT

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, *34*, *3*, 39–48.

Barton, K. E., & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement*, *63*, *4*, 602–614.

Browne, M. W. (1979). The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology*, *32*, 75–86. Doi: 10.1111/j.2044-8317.1979.tb00753.x

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Drasgow, F. (1988). Polychoric and polyserial correlations. In Kotz L, Johnson NL (Eds.), *Encyclopedia of Statistical Sciences*. Vol. 7 (pp. 69–74). New York: Wiley.

Educational Testing Service (2014a). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Educational Testing Service (2014b). *HiSET® technical manual*. Princeton, NJ: Author.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R.L.Linn (Ed). *Educational Measurement*, 3rd Edition, (pp. 105–146). Phoenix, AZ: Oryx Press.

Feldt, L. S., Forsyth, R. A., & Alnot, S. D. (1986). *Iowa Tests of Educational Development, Forms X-8 and Y-8*. Iowa City, IA: University of Iowa.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*. *33*, 613–619.

Furgol, K. Fina, A. and Welch, C. (2011, April). *Establishing validity evidence to assess college readiness through a vertical scale*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hanson, B. A., Zeng, L., & Cui, Z. (2004). STUIRT: *A computer program for IRT scale transformation* [Computer software]. Iowa City, IA: Center for Advanced Study in Measurement and Assessment, University of Iowa.

Holland, P. W., & Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3rd ed.)*. New York, NY: Springer-Verlag.

Lewis. C., & Thayer, D. T. (1986). Unpublished Seminar Notes.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Muthén, B. O. (1998–2004). *Mplus technical appendices*. Los Angeles, CA: Muthén & Muthén.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished Technical Report.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide*. (7th ed.). Los Angeles, CA: Muthén & Muthén.

Olsson, U., Drasgow, F. & Dorans, N.J. (1982). The Polyserial Correlation Coefficient. *Psychometrika*, *47*, 337–347.

Pimentel, S. (2013). *College and Career Readiness Standards for Adult Education*. Washington, DC: U.S. Department of Education, Office of Vocational and Adult Education.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research*, *99*, 323–337.

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Tannenbaum, R. J., & Reese, C. M. (2014). *Recommending passing scores for the high school equivalency test (HiSET®)*. (Research Memorandum 14-06). Princeton, NJ: Educational Testing Service.

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, *23*, 111–136.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states (Research Monograph No. 18)*. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2002). *An analysis of the alignment between mathematics standards and assessments for three states*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20,* 7–25.

Yin, P., Brennan, B., & Kolen, M. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement*, *28*, 274–289.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Skokie, IL: Scientific Software International.

Appendix A: Item Statistics

The tables in Appendix A include the following information:

A. Item Type. MC (multiple choice) or Essay.

B. Item Flag. The item flags are defined as

- A = p-value < 0.20;
- H = p-value > 0.90;
- R = discrimination < 0.25;
- D = distractor chosen by > 20% of high ability test takers; and
- O = omit rate > 5% for MC items and omit rate > 15% for CR items.

C. Observed p-value. Ranging from 0.0 to 1.0.

D. Observed Item-Total Correlation. Ranging from -1.0 to 1.0.

E. Omit Rate. The percentage of test takers omitting the item.

F. a parameter estimates. The IRT parameter used to describe item discrimination.

G. b parameter estimates. The IRT parameter used to describe item difficulty.

H. c parameter estimates. The IRT parameter used to describe the ability to guess the item's correct response.

Note that there are no IRT parameters for the Written Essays.

Table A.1 Item Statistics: Reading Form A

Item Number	Item Type	Item Flag	Observed *p*-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC	R	0.83	0.24	0.04	0.472	-1.605	0.062
2	MC		0.83	0.56	0.04	0.808	-0.594	0.163
3	MC		0.82	0.47	0.03	0.689	-0.500	0.180
4	MC		0.78	0.44	0.11	0.798	-0.351	0.278
5	MC		0.56	0.54	0.11	0.948	-0.173	0.105
6	MC		0.85	0.53	0.09	0.965	-1.144	0.147
7	MC		0.81	0.63	0.08	1.160	-0.332	0.198
8	MC		0.46	0.39	0.29	0.733	1.116	0.246
9	MC		0.89	0.55	0.07	1.870	-0.333	0.270
10	MC		0.84	0.53	0.20	1.760	-0.427	0.121
11	MC		0.84	0.43	0.13	1.790	-0.460	0.143
12	MC	H	0.92	0.56	0.10	1.813	-0.534	0.214
13	MC	H	0.91	0.57	0.14	1.786	0.140	0.211
14	MC		0.63	0.58	0.09	2.064	-0.075	0.312
15	MC		0.68	0.49	0.14	1.907	0.237	0.226
16	MC	R	0.47	0.17	0.07	1.386	0.662	0.294
17	MC		0.66	0.42	0.04	0.737	-0.204	0.248
18	MC		0.70	0.60	0.09	0.992	-0.648	0.239
19	MC		0.58	0.55	0.24	1.555	0.550	0.211
20	MC		0.72	0.35	0.09	0.494	-0.875	0.242
21	MC		0.42	0.48	0.14	1.115	0.893	0.187
22	MC		0.69	0.58	0.14	0.748	-0.561	0.224
23	MC	H	0.90	0.44	0.08	0.847	-0.853	0.271
24	MC		0.71	0.61	0.13	0.952	0.493	0.238
25	MC		0.46	0.60	0.29	1.340	0.752	0.116
26	MC		0.63	0.52	0.07	1.362	0.944	0.247
27	MC		0.80	0.52	0.13	1.048	0.209	0.226
28	MC		0.73	0.43	0.21	1.155	0.104	0.270
29	MC		0.73	0.58	0.08	1.408	0.140	0.224
30	MC		0.64	0.50	0.12	1.294	0.576	0.204
31	MC		0.80	0.56	0.09	1.530	0.138	0.238
32	MC		0.85	0.46	0.07	1.678	-0.246	0.258
33	MC		0.49	0.42	0.51	1.296	0.907	0.276
34	MC	H	0.93	0.66	0.07	1.267	-0.553	0.323
35	MC		0.67	0.60	0.24	0.883	0.410	0.214
36	MC	R	0.34	0.16	0.15	0.882	0.983	0.306
37	MC		0.72	0.50	0.31	1.269	0.316	0.297
38	MC		0.58	0.39	0.24	0.809	0.668	0.241
39	MC		0.44	0.49	0.35	1.691	1.171	0.261
40	MC		0.64	0.47	0.51	0.695	0.612	0.156

Table A.2 Item Statistics: Reading Form B								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	a parameter	b parameter	c parameter
1	MC		0.69	0.48	0.17	1.555	0.514	0.286
2	MC	H	0.94	0.27	0.10	0.856	-0.817	0.157
3	MC		0.77	0.48	0.12	1.124	-0.066	0.156
4	MC	H	0.93	0.57	0.01	1.377	-0.468	0.136
5	MC		0.85	0.37	0.12	0.527	0.308	0.123
6	MC		0.51	0.59	0.04	1.500	0.466	0.236
7	MC		0.87	0.36	0.10	1.416	-0.169	0.172
8	MC		0.82	0.35	0.03	1.207	0.213	0.173
9	MC		0.65	0.36	0.10	1.327	0.867	0.361
10	MC		0.76	0.51	0.06	0.920	-0.027	0.186
11	MC		0.73	0.34	0.09	0.527	-0.267	0.125
12	MC	H	0.93	0.54	0.18	1.322	-0.504	0.221
13	MC		0.82	0.59	0.08	1.480	0.020	0.226
14	MC		0.77	0.64	0.09	1.612	0.096	0.256
15	MC		0.84	0.32	0.03	0.291	-1.199	0.011
16	MC		0.79	0.42	0.11	0.376	-0.620	0.013
17	MC		0.67	0.47	0.03	0.751	-0.179	0.218
18	MC		0.81	0.46	0.03	1.264	0.385	0.498
19	MC		0.43	0.58	0.17	1.235	0.952	0.209
20	MC		0.60	0.59	0.04	1.313	0.071	0.342
21	MC		0.64	0.48	0.12	1.087	0.033	0.209
22	MC		0.89	0.53	0.04	0.963	-0.547	0.239
23	MC		0.67	0.54	0.18	0.816	-0.248	0.192
24	MC		0.28	0.45	0.16	1.205	1.104	0.177
25	MC		0.52	0.51	0.21	0.788	-0.068	0.163
26	MC		0.76	0.54	0.08	0.715	-1.267	0.233
27	MC		0.81	0.54	0.10	1.079	-0.380	0.404
28	MC		0.44	0.49	0.24	1.504	0.676	0.248
29	MC		0.78	0.56	0.09	0.964	-0.413	0.265
30	MC		0.81	0.53	0.09	0.725	-0.768	0.232
31	MC		0.83	0.53	0.06	1.000	-0.523	0.340
32	MC		0.57	0.52	0.22	0.852	0.427	0.193
33	MC	H	0.91	0.45	0.11	0.848	-0.822	0.268
34	MC		0.67	0.43	0.13	1.021	0.277	0.268
35	MC		0.59	0.46	0.09	0.924	-0.295	0.233
36	MC		0.69	0.49	0.19	0.832	0.143	0.342
37	MC		0.69	0.49	0.10	0.854	0.362	0.224
38	MC		0.33	0.34	0.15	0.971	1.798	0.190
39	MC	R	0.40	0.19	0.26	0.509	1.999	0.239
40	MC		0.69	0.56	0.62	1.134	-0.217	0.215

Table A.3 Item Statistics: Reading Form C								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	a parameter	b parameter	c parameter
1	MC		0.84	0.55	0.12	0.894	-1.324	0.248
2	MC		0.81	0.46	0.03	0.688	-1.126	0.239
3	MC		0.77	0.46	0.15	0.558	-0.873	0.256
4	MC	H	0.96	0.53	0.06	0.770	-1.501	0.202
5	MC		0.89	0.43	0.01	0.650	-1.515	0.261
6	MC		0.69	0.43	0.16	0.585	-0.127	0.249
7	MC		0.88	0.54	0.05	0.705	-1.891	0.218
8	MC		0.59	0.61	0.14	0.895	-0.422	0.206
9	MC		0.51	0.51	0.20	0.788	-0.068	0.163
10	MC		0.77	0.55	0.01	0.715	-1.267	0.233
11	MC		0.83	0.53	0.08	1.079	-0.380	0.404
12	MC		0.45	0.50	0.38	1.504	0.676	0.248
13	MC		0.79	0.55	0.05	0.964	-0.413	0.265
14	MC		0.75	0.54	0.12	1.039	-0.361	0.362
15	MC		0.78	0.52	0.05	0.725	-0.768	0.232
16	MC		0.83	0.55	0.12	1.000	-0.523	0.340
17	MC		0.57	0.52	0.10	1.150	0.637	0.247
18	MC		0.68	0.58	0.15	1.349	0.227	0.269
19	MC		0.50	0.46	0.16	0.905	0.477	0.212
20	MC		0.53	0.62	0.10	1.452	0.168	0.243
21	MC		0.39	0.42	0.13	0.808	0.876	0.295
22	MC		0.57	0.50	0.08	0.971	-0.101	0.150
23	MC		0.67	0.59	0.06	1.424	0.107	0.244
24	MC		0.67	0.51	0.05	0.898	0.044	0.197
25	MC	H	0.91	0.58	0.19	2.363	0.227	0.251
26	MC		0.70	0.51	0.06	1.451	0.692	0.228
27	MC		0.78	0.53	0.10	2.084	0.516	0.161
28	MC		0.52	0.53	0.15	2.242	0.742	0.250
29	MC		0.83	0.55	0.09	2.713	0.473	0.285
30	MC		0.36	0.47	0.14	1.835	1.085	0.163
31	MC		0.43	0.47	0.13	2.728	1.125	0.234
32	MC		0.84	0.65	0.12	2.063	0.384	0.266
33	MC		0.85	0.45	0.19	0.982	-0.009	0.322
34	MC		0.34	0.38	0.36	3.363	1.208	0.225
35	MC		0.46	0.46	0.19	1.548	0.859	0.264
36	MC		0.48	0.55	0.28	1.717	0.794	0.212
37	MC		0.65	0.53	0.22	2.301	0.759	0.310
38	MC		0.60	0.56	0.34	1.151	0.485	0.204
39	MC		0.65	0.62	0.26	2.180	0.493	0.224
40	MC		0.72	0.64	0.46	1.514	0.526	0.270

Table A.4 Item Statistics: Writing Form A								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	a parameter	b parameter	c parameter
1	MC		0.81	0.45	0.05	1.045	-0.303	0.324
2	MC		0.83	0.36	0.02	0.644	-1.371	0.079
3	MC		0.82	0.33	0.07	0.614	-0.893	0.194
4	MC		0.47	0.49	0.05	1.682	0.503	0.282
5	MC	R	0.66	0.20	0.08	0.562	-0.389	0.142
6	MC	R	0.45	0.19	0.08	0.527	0.646	0.119
7	MC	H	0.92	0.58	0.15	1.316	-0.715	0.253
8	MC		0.73	0.41	0.15	0.745	-0.512	0.180
9	MC	R	0.58	0.02	0.16	0.391	1.926	0.273
10	MC	H	0.92	0.33	0.07	0.929	-0.969	0.081
11	MC	R	0.54	0.23	0.08	0.642	-0.376	0.106
12	MC	H	0.94	0.47	0.03	1.148	-1.130	0.177
13	MC		0.80	0.42	0.07	1.090	-0.264	0.236
14	MC		0.27	0.34	0.18	1.244	1.371	0.204
15	MC		0.76	0.40	0.10	0.949	-0.240	0.160
16	MC		0.82	0.53	0.15	1.210	-0.352	0.169
17	MC	D	0.25	0.36	0.18	1.482	1.248	0.173
18	MC		0.58	0.46	0.13	1.083	0.335	0.260
19	MC		0.77	0.35	0.11	0.643	-0.294	0.214
20	MC		0.76	0.58	0.05	1.403	-0.258	0.180
21	MC		0.70	0.37	0.03	0.983	-0.182	0.192
22	MC		0.64	0.34	0.11	0.641	0.285	0.155
23	MC		0.88	0.48	0.13	0.956	0.111	0.203
24	MC		0.49	0.54	0.20	1.699	0.691	0.206
25	MC		0.81	0.43	0.21	0.800	-0.323	0.247
26	MC	R	0.42	0.22	0.08	0.989	0.879	0.266
27	MC		0.73	0.54	0.15	1.350	-0.002	0.368
28	MC		0.68	0.38	0.36	0.738	-0.320	0.216
29	MC		0.30	0.38	0.20	1.483	1.270	0.198
30	MC	H	0.91	0.41	0.05	0.765	-1.441	0.227
31	MC		0.49	0.52	0.10	1.393	0.509	0.249
32	MC		0.76	0.44	0.08	1.477	0.084	0.263
33	MC		0.31	0.47	0.44	1.220	1.225	0.152
34	MC		0.52	0.44	0.03	1.073	0.555	0.185
35	MC		0.23	0.37	0.28	1.052	1.007	0.217
36	MC		0.58	0.34	0.18	1.182	0.170	0.320
37	MC		0.55	0.49	0.42	1.889	0.264	0.305
38	MC		0.67	0.44	0.15	1.683	0.393	0.245
39	MC		0.84	0.49	0.11	1.810	-0.159	0.305
40	MC		0.65	0.58	0.16	1.727	-0.044	0.159

Table A.4 Item Statistics: Writing Form A								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.72	0.33	0.11	1.067	0.339	0.291
42	MC		0.68	0.50	0.28	1.280	0.680	0.219
43	MC	R	0.68	0.13	0.07	0.597	-0.076	0.256
44	MC		0.40	0.29	0.08	1.866	1.052	0.286
45	MC		0.56	0.48	0.21	1.316	0.686	0.265
46	MC	R	0.47	0.18	0.10	1.670	0.871	0.283
47	MC		0.34	0.54	0.13	1.326	0.988	0.130
48	MC		0.81	0.49	0.08	1.640	0.137	0.316
49	MC		0.39	0.42	0.29	1.180	0.838	0.185
50	MC		0.81	0.47	0.24	1.001	0.546	0.353
Prompt 1	Essay		0.54	0.45	1.26			
Prompt 2	Essay		0.56	0.46	0.98			

Table A.5 Item Statistics: Writing Form B								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC		0.77	0.42	0.03	0.729	-0.426	0.234
2	MC		0.66	0.53	0.33	1.211	0.029	0.250
3	MC		0.59	0.49	0.10	1.376	0.157	0.229
4	MC		0.88	0.55	0.07	1.306	-0.667	0.262
5	MC		0.56	0.32	0.13	0.529	0.409	0.223
6	MC		0.79	0.37	0.06	1.025	-0.258	0.333
7	MC		0.62	0.30	0.10	0.761	-0.699	0.285
8	MC	H	0.91	0.47	0.03	0.910	-1.350	0.056
9	MC		0.88	0.32	0.06	0.955	-1.430	0.049
10	MC		0.55	0.26	0.10	0.810	1.141	0.380
11	MC		0.77	0.38	0.04	0.880	-0.431	0.173
12	MC		0.76	0.34	0.15	0.695	-0.547	0.097
13	MC		0.67	0.58	0.11	1.727	0.242	0.276
14	MC		0.77	0.41	0.08	1.252	0.139	0.193
15	MC	H	0.90	0.52	0.59	1.144	-1.313	0.228
16	MC		0.71	0.35	0.22	0.860	-0.006	0.318
17	MC		0.65	0.34	0.13	0.645	1.919	0.342
18	MC	H	0.92	0.43	0.06	0.764	-1.409	0.263
19	MC		0.87	0.51	0.08	1.020	-0.739	0.286
20	MC		0.74	0.41	0.63	0.889	-0.373	0.252
21	MC		0.42	0.39	0.08	1.056	0.859	0.202
22	MC		0.52	0.30	0.12	1.766	0.977	0.286
23	MC		0.63	0.44	0.20	1.274	0.247	0.231
24	MC		0.49	0.31	0.26	1.536	0.776	0.334
25	MC		0.83	0.47	0.06	1.099	-0.430	0.233
26	MC		0.43	0.32	0.11	0.657	1.631	0.203
27	MC		0.81	0.55	0.08	1.165	-0.331	0.223
28	MC		0.77	0.44	0.20	1.097	0.053	0.202
29	MC		0.71	0.53	0.40	1.417	0.607	0.239
30	MC		0.59	0.36	0.06	1.325	0.454	0.295
31	MC		0.55	0.37	0.07	1.187	1.010	0.338
32	MC		0.66	0.45	0.15	1.237	0.107	0.188
33	MC		0.25	0.47	0.13	3.530	0.916	0.164
34	MC		0.55	0.38	0.22	1.529	0.533	0.359
35	MC		0.40	0.4	0.07	1.087	0.257	0.177
36	MC		0.58	0.47	0.08	1.317	0.542	0.261
37	MC		0.79	0.28	0.12	0.999	-0.626	0.213
38	MC		0.85	0.46	0.11	1.050	-0.130	0.251
39	MC		0.43	0.28	0.14	0.922	0.806	0.224
40	MC	R	0.72	0.18	0.09	0.626	0.015	0.185

Table A.5 Item Statistics: Writing Form B								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	a parameter	b parameter	c parameter
41	MC		0.22	0.25	0.18	1.167	1.203	0.267
42	MC		0.36	0.42	0.09	1.881	0.982	0.243
43	MC		0.59	0.46	0.18	1.891	0.399	0.236
44	MC		0.47	0.47	0.19	1.026	0.755	0.234
45	MC		0.56	0.41	0.39	1.131	0.473	0.244
46	MC		0.55	0.34	0.30	1.407	0.558	0.387
47	MC		0.56	0.30	0.12	0.741	0.284	0.147
48	MC		0.63	0.34	0.18	0.798	1.028	0.256
49	MC		0.34	0.45	0.25	0.866	0.844	0.183
50	MC		0.63	0.48	0.52	1.176	0.106	0.271
Prompt 1	Essay		0.57	0.46	0.87			
Prompt 2	Essay		0.55	0.45	0.89			
Prompt 3	Essay		0.58	0.47	0.65			

Table A.6 Item Statistics: Writing Form C

Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC		0.55	0.36	0.13	1.250	0.242	0.346
2	MC		0.71	0.30	0.03	1.085	-0.278	0.217
3	MC	H	0.94	0.31	0.02	1.042	-1.159	0.035
4	MC		0.86	0.46	0.06	1.729	-0.021	0.213
5	MC		0.81	0.28	0.04	0.647	-0.315	0.160
6	MC		0.87	0.50	0.07	2.054	0.008	0.285
7	MC		0.62	0.46	0.09	1.582	0.421	0.242
8	MC	H	0.90	0.44	0.07	1.798	0.043	0.240
9	MC	D	0.29	0.28	0.11	0.583	2.130	0.149
10	MC		0.86	0.51	0.04	2.241	-0.056	0.289
11	MC		0.53	0.34	0.18	1.674	-0.008	0.257
12	MC	R	0.58	0.23	0.13	0.759	0.659	0.278
13	MC		0.74	0.43	0.04	0.961	0.049	0.318
14	MC		0.87	0.55	0.03	1.370	0.093	0.267
15	MC		0.58	0.40	0.08	0.805	0.096	0.228
16	MC		0.64	0.43	0.36	0.816	-0.313	0.205
17	MC		0.39	0.43	0.03	1.565	0.694	0.236
18	MC		0.65	0.43	0.13	1.183	0.418	0.252
19	MC		0.57	0.55	0.04	1.170	0.017	0.144
20	MC		0.76	0.40	0.13	0.898	-0.296	0.352
21	MC		0.89	0.52	0.01	1.467	-0.659	0.318
22	MC		0.64	0.55	0.07	1.843	0.355	0.190
23	MC		0.86	0.42	0.31	1.327	-0.490	0.225
24	MC		0.46	0.30	0.22	0.742	0.518	0.161
25	MC		0.37	0.47	0.04	1.008	0.744	0.229
26	MC		0.69	0.41	0.06	0.776	-0.388	0.203
27	MC		0.53	0.46	0.83	1.106	-0.186	0.246
28	MC		0.49	0.41	0.14	0.872	0.662	0.182
29	MC		0.45	0.41	0.14	1.234	0.536	0.280
30	MC		0.59	0.39	0.07	0.794	-0.025	0.150
31	MC	H	0.94	0.50	0.07	1.177	-1.215	0.244
32	MC		0.50	0.47	0.21	0.856	0.579	0.245
33	MC		0.42	0.44	0.13	2.036	0.524	0.270
34	MC		0.65	0.41	0.09	0.719	-0.723	0.094
35	MC		0.75	0.54	0.23	1.479	0.257	0.273
36	MC		0.53	0.39	0.10	0.834	0.577	0.190
37	MC		0.43	0.40	0.11	0.871	0.537	0.276
38	MC	R	0.59	0.13	0.19	0.444	0.614	0.340
39	MC		0.40	0.41	0.47	1.923	1.115	0.154
40	MC		0.73	0.55	0.03	1.913	0.141	0.276

Table A.6 Item Statistics: Writing Form C

Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.54	0.45	0.11	1.144	0.418	0.217
42	MC		0.86	0.62	0.03	1.629	-0.520	0.243
43	MC		0.55	0.35	0.37	0.823	0.632	0.211
44	MC		0.61	0.37	0.10	0.881	0.086	0.281
45	MC		0.68	0.53	0.07	1.404	0.008	0.308
46	MC	R	0.52	0.18	0.07	0.947	1.124	0.188
47	MC		0.36	0.40	0.52	0.849	0.910	0.244
48	MC		0.49	0.46	0.22	1.144	0.413	0.247
49	MC		0.45	0.54	0.13	1.086	0.439	0.106
50	MC		0.59	0.35	0.16	0.676	1.157	0.221
Prompt 1	Essay		0.57	0.49	0.61			
Prompt 2	Essay		0.57	0.46	0.76			
Prompt 3	Essay		0.57	0.42	0.46			

Table A.7 Item Statistics: Mathematics Form A

Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC		0.42	0.55	1.40	1.031	-0.561	0.150
2	MC		0.55	0.44	0.45	1.008	-0.285	0.354
3	MC		0.58	0.54	0.41	0.975	-0.374	0.165
4	MC		0.89	0.56	0.15	1.057	-0.174	0.130
5	MC		0.59	0.61	0.30	1.171	-0.029	0.140
6	MC		0.61	0.61	0.10	1.099	-0.400	0.098
7	MC		0.28	0.38	0.92	1.756	0.348	0.271
8	MC		0.26	0.46	0.69	0.844	0.238	0.180
9	MC		0.55	0.28	0.53	0.814	0.532	0.379
10	MC		0.36	0.47	0.20	0.996	0.441	0.164
11	MC		0.81	0.43	0.06	0.804	-0.245	0.125
12	MC		0.72	0.35	0.10	0.910	0.378	0.094
13	MC	A	0.16	0.44	0.50	1.309	0.284	0.114
14	MC		0.52	0.51	0.15	0.768	0.563	0.221
15	MC		0.59	0.42	0.34	0.817	1.281	0.230
16	MC		0.31	0.29	1.28	0.820	1.447	0.164
17	MC		0.21	0.29	1.33	0.879	1.450	0.188
18	MC		0.38	0.66	0.07	1.753	0.415	0.175
19	MC	A	0.19	0.29	0.40	1.358	2.260	0.156
20	MC		0.23	0.25	1.32	1.420	0.950	0.202
21	MC		0.68	0.49	0.10	0.730	-0.100	0.112
22	MC		0.23	0.37	0.90	1.149	0.903	0.153
23	MC		0.27	0.51	0.29	1.257	0.940	0.203
24	MC	AR	0.19	0.20	1.05	0.923	1.018	0.142
25	MC	AD	0.09	0.32	0.75	1.713	1.190	0.096
26	MC	ARD	0.18	0.18	1.13	1.419	1.143	0.106
27	MC		0.46	0.31	0.28	1.662	0.973	0.242
28	MC		0.68	0.55	0.18	0.743	-0.702	0.132
29	MC		0.22	0.35	0.22	0.657	1.627	0.155
30	MC	A	0.18	0.27	0.61	0.787	1.594	0.152
31	MC		0.74	0.45	0.05	0.668	-0.894	0.034
32	MC		0.20	0.50	0.24	1.641	0.808	0.118
33	MC		0.36	0.51	0.10	1.334	1.118	0.148
34	MC		0.25	0.43	0.16	0.865	0.987	0.107
35	MC	R	0.24	0.18	0.81	0.671	2.080	0.130
36	MC	R	0.33	0.10	0.17	1.049	1.881	0.190
37	MC	R	0.27	0.20	0.79	0.737	1.850	0.127
38	MC	R	0.35	0.24	0.95	0.621	2.233	0.157
39	MC	ARD	0.16	0.20	0.52	0.987	1.805	0.198
40	MC	RD	0.21	0.17	0.17	1.106	2.185	0.210

Table A.7 Item Statistics: Mathematics Form A

Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.21	0.30	0.63	1.476	1.485	0.178
42	MC	ARD	0.16	0.16	0.40	1.398	1.803	0.216
43	MC	ARD	0.07	0.21	0.93	1.493	1.354	0.087
44	MC		0.36	0.43	0.32	0.615	2.555	0.157
45	MC	AD	0.16	0.25	0.68	1.656	1.638	0.151
46	MC	AD	0.14	0.28	0.85	2.501	1.470	0.130
47	MC		0.21	0.35	0.43	1.262	1.620	0.102
48	MC		0.71	0.44	0.32	1.572	0.566	0.170
49	MC	R	0.24	0.18	0.57	1.242	1.681	0.137
50	MC	AD	0.10	0.29	0.73	1.161	1.534	0.043

Table A.8 Item Statistics: Mathematics Form B								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC		0.62	0.56	0.44	0.710	-1.670	0.248
2	MC		0.85	0.57	0.22	0.908	-1.076	0.187
3	MC		0.47	0.61	0.21	1.322	-0.026	0.135
4	MC		0.51	0.57	1.27	0.687	0.123	0.162
5	MC		0.36	0.50	0.08	1.035	0.814	0.177
6	MC		0.25	0.47	0.87	0.616	0.353	0.142
7	MC		0.71	0.55	0.41	1.546	0.274	0.223
8	MC		0.54	0.59	0.21	1.127	0.207	0.167
9	MC		0.55	0.34	0.14	0.719	0.365	0.214
10	MC		0.47	0.35	0.42	0.648	0.583	0.178
11	MC	A	0.12	0.54	0.87	1.235	0.654	0.130
12	MC		0.27	0.34	0.53	0.763	0.789	0.179
13	MC		0.21	0.41	0.78	0.940	1.067	0.144
14	MC		0.40	0.37	0.70	0.961	0.844	0.147
15	MC		0.49	0.44	0.26	1.014	0.533	0.122
16	MC	AD	0.04	0.55	0.16	1.500	1.576	0.069
17	MC		0.24	0.50	0.27	0.814	0.843	0.158
18	MC	R	0.29	0.20	1.13	1.097	1.062	0.208
19	MC		0.35	0.36	1.22	1.670	1.113	0.244
20	MC		0.20	0.48	0.68	1.318	0.720	0.110
21	MC	A	0.17	0.37	0.69	0.848	1.022	0.211
22	MC		0.28	0.65	0.31	1.264	1.098	0.157
23	MC	AD	0.13	0.44	0.37	2.321	1.093	0.187
24	MC		0.81	0.56	0.13	0.927	-0.573	0.152
25	MC	R	0.37	0.24	0.62	0.764	1.142	0.187
26	MC	R	0.28	0.23	0.70	1.124	1.332	0.214
27	MC	ARD	0.04	0.23	1.68	1.042	1.063	0.101
28	MC		0.23	0.32	0.60	0.910	1.427	0.190
29	MC	R	0.26	0.24	0.37	0.992	1.178	0.177
30	MC		0.21	0.26	0.30	0.498	1.966	0.175
31	MC		0.45	0.47	0.36	0.698	1.218	0.143
32	MC	R	0.24	0.17	1.29	1.390	1.345	0.157
33	MC		0.22	0.61	0.23	1.619	1.019	0.104
34	MC	RD	0.29	0.23	0.23	0.698	1.871	0.129
35	MC		0.33	0.40	0.52	0.705	1.358	0.194
36	MC		0.29	0.33	0.48	0.838	1.999	0.203
37	MC		0.26	0.32	0.36	1.000	1.308	0.078
38	MC	R	0.22	0.21	1.10	0.799	2.051	0.148
39	MC	ARD	0.16	0.23	0.43	1.701	1.291	0.159
40	MC	ARD	0.13	0.09	0.82	1.222	1.764	0.156

Table A.8 Item Statistics: Mathematics Form B								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC	RD	0.23	0.21	0.70	1.334	1.636	0.189
42	MC	AD	0.14	0.33	0.32	1.507	1.658	0.150
43	MC	RD	0.21	0.17	0.50	1.491	2.012	0.214
44	MC	ARD	0.19	0.11	0.48	1.727	2.059	0.200
45	MC	RD	0.22	0.18	1.09	0.625	2.700	0.171
46	MC	ARD	0.15	0.20	0.72	1.112	1.707	0.087
47	MC		0.23	0.32	0.57	0.749	1.987	0.083
48	MC		0.25	0.25	0.97	1.081	2.115	0.134
49	MC	AD	0.15	0.32	0.55	1.132	1.901	0.122
50	MC	AD	0.09	0.39	0.91	1.338	1.647	0.069

Table A.9 Item Statistics: Mathematics Form C								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC		0.72	0.65	0.18	0.981	-1.359	0.036
2	MC		0.75	0.62	0.24	1.346	-0.520	0.126
3	MC		0.70	0.42	0.17	0.996	-0.487	0.158
4	MC		0.29	0.32	1.74	1.235	0.262	0.186
5	MC		0.41	0.42	0.59	1.082	0.596	0.222
6	MC		0.64	0.51	0.10	0.903	-0.024	0.176
7	MC		0.90	0.32	0.11	0.594	-1.594	0.042
8	MC	AD	0.11	0.34	0.33	1.826	1.634	0.136
9	MC		0.22	0.51	0.63	1.470	0.392	0.101
10	MC		0.64	0.49	0.19	0.878	0.149	0.082
11	MC		0.37	0.45	1.11	1.194	0.614	0.127
12	MC		0.26	0.26	0.97	0.725	1.238	0.277
13	MC		0.54	0.44	0.97	0.824	0.735	0.231
14	MC		0.36	0.51	0.81	1.130	0.875	0.187
15	MC		0.29	0.25	1.50	1.259	1.213	0.252
16	MC	AD	0.18	0.29	0.60	0.991	0.925	0.126
17	MC		0.20	0.37	1.08	1.234	1.093	0.161
18	MC		0.38	0.55	0.18	1.029	0.703	0.188
19	MC		0.76	0.59	0.23	0.718	0.344	0.217
20	MC	RD	0.24	0.21	0.41	2.368	1.764	0.209
21	MC		0.31	0.46	0.19	1.359	0.840	0.114
22	MC		0.34	0.28	0.66	1.473	1.654	0.306
23	MC		0.20	0.50	0.64	1.119	0.751	0.120
24	MC		0.27	0.36	0.71	0.783	1.348	0.177
25	MC	ARD	0.13	0.21	0.62	1.144	0.972	0.110
26	MC	RD	0.20	0.12	1.07	1.128	1.692	0.238
27	MC		0.75	0.45	0.08	0.668	-0.894	0.034
28	MC		0.46	0.44	0.15	1.586	0.789	0.133
29	MC		0.30	0.53	0.29	1.285	0.646	0.196
30	MC	AD	0.05	0.35	0.38	1.326	1.746	0.124
31	MC	AD	0.05	0.34	0.47	1.523	1.820	0.136
32	MC		0.42	0.30	0.12	0.918	1.779	0.318
33	MC	AR	0.20	0.24	0.75	0.804	1.496	0.121
34	MC	ARD	0.19	0.22	0.62	1.118	1.293	0.118
35	MC	RD	0.25	0.20	0.56	0.765	2.086	0.220
36	MC	AD	0.10	0.30	0.86	1.695	1.248	0.131
37	MC		0.33	0.35	0.62	1.926	1.185	0.214
38	MC	R	0.23	0.23	0.78	0.815	1.710	0.128
39	MC	A	0.20	0.25	0.90	0.802	2.836	0.140
40	MC	AD	0.10	0.35	1.29	1.120	1.790	0.182

Table A.9 Item Statistics: Mathematics Form C								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.24	0.29	0.55	0.886	1.360	0.108
42	MC		0.47	0.45	0.35	1.358	0.887	0.242
43	MC	AD	0.14	0.25	0.47	1.206	2.010	0.158
44	MC	ARD	0.09	0.17	0.81	1.210	1.822	0.124
45	MC	R	0.34	0.12	0.60	1.178	1.492	0.155
46	MC	A	0.18	0.37	0.95	1.119	2.017	0.154
47	MC	AR	0.17	0.23	0.86	0.954	2.631	0.147
48	MC	ARD	0.09	0.12	1.01	0.989	2.484	0.134
49	MC	AD	0.10	0.34	0.63	1.342	1.464	0.061
50	MC	AD	0.05	0.38	0.54	1.401	1.494	0.044

Table A.10 Item Statistics: Science Form A

Item Number	Item Type	Item Flag	Observed *p*-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC		0.55	0.37	0.10	1.205	0.883	0.270
2	MC	R	0.47	0.21	0.10	0.972	1.499	0.279
3	MC		0.50	0.44	0.24	1.103	0.637	0.225
4	MC		0.39	0.48	0.25	1.489	1.104	0.236
5	MC		0.52	0.44	0.11	0.927	0.623	0.130
6	MC		0.34	0.34	0.31	0.910	1.895	0.276
7	MC		0.76	0.63	0.13	1.061	0.265	0.254
8	MC		0.49	0.52	0.14	1.488	1.191	0.172
9	MC		0.53	0.40	0.28	0.892	1.121	0.267
10	MC		0.74	0.51	0.11	1.074	0.417	0.164
11	MC	H	0.91	0.54	0.20	0.854	-0.013	0.257
12	MC	H	0.96	0.51	0.06	1.190	-0.860	0.230
13	MC	H	0.96	0.49	0.06	0.900	-1.090	0.210
14	MC		0.59	0.71	0.14	1.870	0.450	0.200
15	MC	RD	0.26	0.13	0.11	1.110	1.440	0.220
16	MC		0.78	0.51	0.14	1.470	0.280	0.180
17	MC		0.77	0.58	0.08	1.230	-0.150	0.230
18	MC		0.48	0.48	0.20	1.900	1.030	0.200
19	MC		0.60	0.55	0.07	1.130	0.420	0.260
20	MC		0.69	0.48	0.06	0.930	0.560	0.270
21	MC		0.75	0.59	0.07	1.453	-0.129	0.148
22	MC	H	0.91	0.52	0.10	1.324	-0.332	0.124
23	MC		0.80	0.53	0.10	1.218	-0.204	0.091
24	MC		0.75	0.62	0.17	1.226	0.233	0.207
25	MC		0.65	0.48	0.22	1.135	0.164	0.211
26	MC	H	0.90	0.61	0.11	1.499	-0.119	0.246
27	MC		0.81	0.45	0.14	0.734	-1.446	0.243
28	MC	RD	0.26	0.21	0.29	0.734	2.173	0.229
29	MC		0.50	0.51	0.20	1.534	0.639	0.223
30	MC		0.37	0.32	0.18	1.927	1.130	0.205
31	MC		0.53	0.36	0.22	1.593	0.757	0.282
32	MC		0.29	0.43	0.34	1.645	1.409	0.259
33	MC		0.75	0.41	0.25	1.252	0.344	0.325
34	MC		0.59	0.50	0.19	2.112	0.711	0.285
35	MC	R	0.45	0.24	0.24	0.735	1.636	0.296
36	MC		0.44	0.63	0.24	1.880	0.910	0.190
37	MC		0.90	0.36	0.12	0.880	0.060	0.210
38	MC		0.46	0.46	0.13	1.290	0.870	0.250
39	MC	A	0.19	0.43	0.17	1.990	1.510	0.230
40	MC		0.73	0.46	0.19	1.060	0.390	0.220

Table A.10 Item Statistics: Science Form A

Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.52	0.30	0.13	0.840	1.270	0.180
42	MC	R	0.34	0.21	0.28	0.840	1.570	0.190
43	MC		0.74	0.66	0.24	2.191	0.319	0.174
44	MC		0.59	0.57	0.23	2.138	0.725	0.268
45	MC		0.41	0.34	0.17	1.118	1.155	0.152
46	MC		0.49	0.26	0.13	0.820	0.732	0.171
47	MC		0.74	0.54	0.18	1.814	0.359	0.209
48	MC		0.54	0.37	0.20	1.159	1.038	0.266
49	MC		0.53	0.56	0.28	2.250	0.805	0.252
50	MC		0.61	0.32	0.32	1.126	0.960	0.188

Table A.11 Item Statistics: Science Form B								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	a parameter	b parameter	c parameter
1	MC		0.73	0.59	0.02	1.480	-0.007	0.274
2	MC		0.84	0.48	0.02	0.705	-0.334	0.040
3	MC		0.76	0.47	0.25	1.125	0.060	0.163
4	MC		0.52	0.49	0.48	2.193	1.006	0.236
5	MC		0.33	0.49	0.08	0.917	1.067	0.162
6	MC		0.60	0.41	0.23	1.244	0.642	0.300
7	MC	R	0.79	0.21	0.13	0.607	-0.397	0.012
8	MC	AD	0.18	0.30	0.14	1.329	1.462	0.183
9	MC		0.59	0.52	0.26	2.142	0.785	0.284
10	MC		0.78	0.54	0.10	2.320	0.641	0.263
11	MC		0.77	0.56	0.19	2.008	0.663	0.295
12	MC		0.54	0.48	0.12	1.062	0.844	0.283
13	MC		0.60	0.35	0.29	1.450	1.034	0.341
14	MC		0.41	0.41	0.30	0.972	1.020	0.181
15	MC		0.73	0.55	0.11	0.932	0.279	0.198
16	MC		0.51	0.51	0.06	1.190	0.725	0.193
17	MC		0.78	0.57	0.11	1.025	0.317	0.268
18	MC		0.68	0.66	0.23	1.282	0.362	0.156
19	MC		0.76	0.69	0.12	1.442	0.383	0.125
20	MC		0.32	0.55	0.10	1.772	1.025	0.197
21	MC		0.70	0.50	0.11	1.049	0.721	0.134
22	MC		0.76	0.51	0.15	0.965	-0.249	0.184
23	MC		0.48	0.60	0.13	2.184	0.849	0.201
24	MC		0.80	0.41	0.21	0.639	0.120	0.224
25	MC		0.34	0.26	0.19	1.082	1.192	0.319
26	MC		0.28	0.34	0.25	1.581	1.829	0.252
27	MC	R	0.51	0.13	0.32	0.520	1.059	0.241
28	MC		0.64	0.42	0.10	1.844	0.770	0.302
29	MC		0.79	0.54	0.08	1.474	0.203	0.176
30	MC		0.69	0.50	0.08	1.337	0.134	0.222
31	MC	R	0.36	0.20	0.08	0.303	2.463	0.086
32	MC		0.62	0.34	0.07	1.262	0.459	0.281
33	MC		0.34	0.33	0.11	1.504	1.737	0.233
34	MC		0.56	0.65	0.12	2.100	0.826	0.212
35	MC		0.50	0.61	0.19	2.810	1.050	0.220
36	MC		0.41	0.64	0.07	2.560	1.180	0.190
37	MC		0.56	0.41	0.04	1.000	0.810	0.260
38	MC		0.68	0.34	0.12	1.070	0.570	0.230
39	MC		0.57	0.40	0.06	0.830	1.050	0.210
40	MC		0.56	0.45	0.26	1.940	1.130	0.280

Table A.11 Item Statistics: Science Form B								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.78	0.54	0.11	1.190	0.150	0.280
42	MC		0.71	0.43	0.06	1.150	0.630	0.300
43	MC		0.36	0.41	0.08	0.891	1.103	0.240
44	MC	ARD	0.18	0.23	0.10	1.687	1.591	0.167
45	MC		0.35	0.37	0.23	1.326	1.255	0.183
46	MC		0.45	0.53	0.23	1.066	1.007	0.159
47	MC		0.68	0.53	0.15	0.892	0.675	0.246
48	MC	R	0.23	0.15	0.18	2.194	1.688	0.187
49	MC		0.83	0.47	0.12	0.923	0.063	0.267
50	MC		0.82	0.52	0.32	0.938	0.277	0.174

Table A.12 Item Statistics: Science Form C								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC		0.70	0.55	0.10	0.855	-0.416	0.034
2	MC		0.79	0.30	0.15	0.768	-0.504	0.101
3	MC	R	0.54	0.20	0.10	0.284	0.991	0.079
4	MC		0.64	0.48	0.09	0.732	0.170	0.187
5	MC	R	0.47	0.14	0.49	0.468	1.333	0.209
6	MC		0.77	0.55	0.27	0.699	-0.299	0.291
7	MC		0.66	0.57	0.23	1.203	0.150	0.223
8	MC		0.59	0.35	0.18	1.418	0.842	0.298
9	MC		0.74	0.52	0.19	1.079	0.497	0.228
10	MC		0.87	0.53	0.05	0.818	-0.290	0.242
11	MC		0.52	0.39	0.23	0.722	0.728	0.346
12	MC		0.71	0.53	0.12	1.235	0.484	0.331
13	MC		0.78	0.49	0.18	1.102	0.337	0.291
14	MC		0.66	0.63	0.10	1.290	0.352	0.185
15	MC		0.77	0.49	0.05	1.157	0.087	0.256
16	MC		0.44	0.39	0.23	0.871	1.038	0.175
17	MC		0.75	0.67	0.05	1.980	0.520	0.310
18	MC		0.42	0.51	0.23	1.550	1.040	0.170
19	MC		0.84	0.63	0.15	1.250	-0.120	0.290
20	MC	R	0.44	0.24	0.06	0.730	0.980	0.260
21	MC		0.84	0.56	0.09	1.210	0.340	0.260
22	MC		0.85	0.37	0.09	0.950	-0.100	0.190
23	MC		0.72	0.52	0.16	1.340	0.180	0.200
24	MC	R	0.84	0.18	0.05	0.690	0.340	0.260
25	MC		0.81	0.40	0.26	0.955	0.771	0.382
26	MC	H	0.90	0.39	0.13	1.075	-0.888	0.222
27	MC		0.57	0.50	0.05	1.203	0.631	0.200
28	MC		0.80	0.50	0.17	1.139	-0.003	0.223
29	MC		0.42	0.49	0.21	1.639	1.020	0.225
30	MC		0.44	0.45	0.40	1.361	1.300	0.261
31	MC		0.49	0.27	0.35	0.822	1.619	0.260
32	MC		0.57	0.25	0.35	0.644	1.946	0.299
33	MC		0.54	0.58	0.29	1.576	0.593	0.191
34	MC	H	0.92	0.66	0.10	0.981	-0.503	0.218
35	MC	R	0.29	0.01	0.29	1.617	2.069	0.212
36	MC		0.78	0.56	0.21	0.969	0.782	0.247
37	MC		0.36	0.45	0.46	1.269	1.012	0.170
38	MC		0.62	0.54	0.10	0.762	1.515	0.188
39	MC		0.43	0.60	0.23	2.551	0.863	0.213
40	MC		0.57	0.45	0.21	1.642	1.029	0.174

Table A.12 Item Statistics: Science Form C								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.42	0.48	0.18	1.790	1.104	0.246
42	MC		0.63	0.61	0.22	1.754	0.827	0.311
43	MC		0.60	0.57	0.21	2.207	0.845	0.213
44	MC		0.48	0.40	0.28	1.714	1.396	0.244
45	MC		0.49	0.42	0.18	2.378	1.136	0.200
46	MC		0.63	0.60	0.27	2.144	0.833	0.314
47	MC		0.36	0.45	0.24	1.112	1.397	0.124
48	MC		0.51	0.55	0.26	1.761	1.094	0.254
49	MC		0.54	0.44	0.46	1.282	1.221	0.229
50	MC		0.34	0.52	0.81	1.730	1.172	0.182

Table A.13 Item Statistics: Social Studies Form A								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	a parameter	b parameter	c parameter
1	MC		0.82	0.38	0.09	0.588	-1.267	0.059
2	MC		0.64	0.46	0.20	0.836	-0.174	0.280
3	MC		0.66	0.58	0.11	0.945	-0.251	0.233
4	MC		0.56	0.33	0.11	0.696	0.609	0.303
5	MC		0.49	0.64	0.17	1.070	-0.151	0.162
6	MC		0.56	0.51	0.23	1.292	0.339	0.296
7	MC	H	0.99	0.31	0.01	0.941	-1.308	0.252
8	MC		0.88	0.35	0.04	0.809	-0.605	0.279
9	MC	RH	0.93	0.22	0.04	0.781	-1.161	0.238
10	MC		0.64	0.63	0.10	1.558	0.242	0.390
11	MC		0.75	0.48	0.17	1.227	0.680	0.257
12	MC		0.58	0.40	0.45	1.664	0.503	0.246
13	MC		0.68	0.58	0.07	0.843	-0.700	0.153
14	MC		0.56	0.22	0.09	0.522	-0.051	0.172
15	MC		0.74	0.51	0.08	0.674	-0.899	0.120
16	MC		0.58	0.32	0.18	0.633	0.847	0.265
17	MC		0.61	0.47	0.11	1.348	0.291	0.236
18	MC	H	0.91	0.32	0.07	0.745	-0.451	0.284
19	MC		0.64	0.51	0.09	1.240	0.474	0.289
20	MC		0.37	0.48	0.11	1.333	1.002	0.313
21	MC		0.47	0.29	0.14	0.721	0.159	0.223
22	MC	R	0.86	-0.04	0.07	0.260	-0.921	0.343
23	MC	R	0.47	-0.01	0.21	0.739	1.860	0.334
24	MC		0.87	0.53	0.03	0.924	-0.596	0.245
25	MC		0.63	0.28	0.11	0.626	-0.514	0.277
26	MC		0.78	0.42	0.12	0.689	-1.180	0.290
27	MC	RD	0.27	0.14	0.13	0.680	1.899	0.241
28	MC		0.59	0.30	0.10	0.780	1.223	0.259
29	MC		0.69	0.36	0.12	1.480	0.272	0.332
30	MC		0.25	0.39	0.16	1.355	1.328	0.180
31	MC		0.71	0.46	0.08	1.078	0.058	0.178
32	MC		0.55	0.44	0.10	0.970	0.218	0.229
33	MC		0.46	0.57	0.12	1.063	0.553	0.140
34	MC		0.53	0.60	0.09	1.808	0.630	0.260
35	MC		0.64	0.43	0.11	1.082	0.469	0.307
36	MC		0.42	0.45	0.09	1.154	0.811	0.275
37	MC		0.55	0.53	0.07	1.487	0.696	0.315
38	MC	D	0.27	0.35	0.07	1.130	1.959	0.265
39	MC		0.23	0.46	0.09	1.424	1.128	0.184
40	MC		0.62	0.55	0.10	1.240	0.236	0.189

Table A.13 Item Statistics: Social Studies Form A

Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.58	0.42	0.13	0.915	0.775	0.287
42	MC		0.56	0.41	0.21	1.057	0.322	0.237
43	MC		0.41	0.57	0.10	1.440	0.585	0.249
44	MC		0.58	0.35	0.18	1.339	0.583	0.255
45	MC		0.42	0.48	0.20	1.860	1.060	0.201
46	MC		0.48	0.40	0.16	1.819	0.827	0.292
47	MC		0.58	0.46	0.17	1.872	0.884	0.204
48	MC		0.61	0.31	0.28	1.335	0.781	0.188
49	MC		0.51	0.49	0.27	2.427	1.005	0.277
50	MC		0.44	0.43	0.50	2.350	0.942	0.285

Table A.14 Item Statistics: Social Studies Form B								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC		0.68	0.47	0.29	1.054	0.018	0.167
2	MC	R	0.84	0.20	0.07	0.764	-0.675	0.018
3	MC		0.84	0.58	0.03	1.321	-0.455	0.297
4	MC		0.45	0.47	0.17	1.418	0.576	0.190
5	MC		0.45	0.43	0.18	1.116	0.837	0.247
6	MC		0.58	0.51	0.12	0.769	0.304	0.131
7	MC	A	0.16	0.55	0.03	2.464	1.524	0.095
8	MC		0.62	0.64	0.11	1.415	-0.162	0.173
9	MC	H	0.94	0.56	0.01	1.612	-0.976	0.038
10	MC		0.78	0.64	0.02	1.252	-0.359	0.108
11	MC		0.72	0.59	0.07	1.186	-0.649	0.049
12	MC		0.46	0.43	0.09	0.854	0.284	0.094
13	MC		0.78	0.47	0.10	0.842	-0.317	0.033
14	MC		0.64	0.59	0.08	1.227	0.283	0.263
15	MC		0.82	0.30	0.07	0.822	-0.252	0.263
16	MC		0.64	0.47	0.09	1.021	0.411	0.315
17	MC		0.53	0.43	0.07	1.158	0.762	0.221
18	MC		0.80	0.58	0.07	1.343	-0.121	0.273
19	MC		0.55	0.51	0.18	1.137	0.041	0.269
20	MC		0.67	0.56	0.03	1.154	0.019	0.174
21	MC		0.41	0.27	0.08	1.291	0.688	0.279
22	MC		0.81	0.43	0.03	0.731	-0.238	0.244
23	MC		0.82	0.26	0.06	0.782	-0.163	0.223
24	MC	R	0.88	0.22	0.01	0.388	-1.320	0.281
25	MC		0.60	0.46	0.07	1.191	0.578	0.286
26	MC		0.51	0.48	0.09	0.981	0.709	0.398
27	MC		0.41	0.45	0.11	0.922	0.696	0.272
28	MC		0.28	0.33	0.18	1.413	1.091	0.207
29	MC		0.58	0.50	0.38	1.054	0.403	0.230
30	MC		0.50	0.50	0.20	0.916	0.460	0.195
31	MC	R	0.46	0.23	0.04	1.570	1.053	0.264
32	MC		0.60	0.43	0.16	1.108	0.963	0.414
33	MC		0.57	0.50	0.15	1.018	0.339	0.282
34	MC		0.39	0.49	0.13	2.138	0.635	0.269
35	MC		0.33	0.40	0.07	0.901	0.881	0.167
36	MC		0.66	0.60	0.18	1.416	-0.222	0.192
37	MC		0.69	0.36	0.02	0.683	-1.001	0.152
38	MC	R	0.78	0.22	0.04	0.748	-0.548	0.112
39	MC		0.74	0.37	0.01	0.530	-0.334	0.134
40	MC		0.38	0.33	0.12	0.893	0.807	0.195

Table A.14 Item Statistics: Social Studies Form B								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.63	0.49	0.17	0.886	0.064	0.258
42	MC		0.48	0.41	0.09	0.671	1.147	0.252
43	MC		0.75	0.45	0.04	1.003	0.041	0.358
44	MC		0.69	0.48	0.15	1.347	0.387	0.251
45	MC		0.57	0.37	0.10	1.210	0.615	0.203
46	MC		0.66	0.38	0.11	0.746	0.554	0.198
47	MC		0.58	0.36	0.15	1.021	0.695	0.208
48	MC		0.64	0.52	0.09	1.171	0.865	0.168
49	MC		0.62	0.52	0.17	1.690	0.611	0.220
50	MC		0.53	0.32	0.30	1.156	1.214	0.240

Table A.15 Item Statistics: Social Studies Form C								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
1	MC		0.77	0.49	0.13	1.172	-0.204	0.152
2	MC		0.68	0.52	0.16	1.107	0.183	0.202
3	MC		0.75	0.53	0.16	1.299	0.157	0.178
4	MC		0.50	0.41	0.20	1.064	0.801	0.283
5	MC		0.45	0.38	0.38	1.011	0.604	0.312
6	MC		0.63	0.63	0.16	1.415	-0.162	0.173
7	MC	H	0.96	0.58	0.01	1.612	-0.976	0.038
8	MC	R	0.71	0.19	0.08	0.635	-0.720	0.309
9	MC		0.52	0.41	0.21	1.596	0.489	0.309
10	MC		0.68	0.40	0.15	0.961	0.343	0.144
11	MC		0.62	0.47	0.16	1.013	-0.460	0.282
12	MC		0.59	0.58	0.22	1.598	0.488	0.306
13	MC		0.48	0.54	0.16	1.208	0.597	0.176
14	MC		0.76	0.53	0.11	1.205	-0.379	0.195
15	MC		0.39	0.25	0.23	0.838	1.387	0.316
16	MC		0.46	0.42	0.03	0.854	0.284	0.094
17	MC		0.78	0.45	0.15	0.842	-0.317	0.033
18	MC		0.67	0.60	0.11	1.227	0.283	0.263
19	MC		0.43	0.47	0.35	1.513	0.840	0.172
20	MC		0.53	0.56	0.32	1.841	0.729	0.265
21	MC		0.76	0.37	0.09	1.000	-0.178	0.175
22	MC		0.66	0.51	0.16	0.799	0.056	0.097
23	MC		0.50	0.30	0.10	0.668	1.001	0.251
24	MC		0.55	0.53	0.03	0.847	-0.532	0.131
25	MC		0.60	0.42	0.04	1.834	0.837	0.350
26	MC		0.59	0.32	0.25	0.633	0.847	0.265
27	MC		0.60	0.48	0.11	1.348	0.291	0.236
28	MC	H	0.90	0.35	0.09	0.745	-0.451	0.284
29	MC		0.63	0.52	0.10	1.240	0.474	0.289
30	MC		0.27	0.46	0.09	1.913	1.116	0.176
31	MC		0.34	0.40	0.12	1.597	1.200	0.247
32	MC		0.51	0.34	0.32	0.739	0.777	0.287
33	MC		0.61	0.45	0.18	1.043	0.570	0.255
34	MC		0.74	0.49	0.19	1.411	0.201	0.330
35	MC	R	0.44	0.20	0.10	0.783	0.984	0.237
36	MC	R	0.31	0.15	0.67	1.337	1.417	0.245
37	MC	R	0.86	0.24	0.04	0.388	-1.320	0.281
38	MC		0.59	0.48	0.06	1.191	0.578	0.286
39	MC		0.54	0.52	0.07	0.981	0.709	0.398
40	MC		0.43	0.50	0.09	0.922	0.696	0.272

Table A.15 Item Statistics: Social Studies Form C								
Item Number	Item Type	Item Flag	Observed p-value	Observed Item-Total Correlation	Omit Rate	*a* parameter	*b* parameter	*c* parameter
41	MC		0.44	0.49	0.36	1.755	0.772	0.219
42	MC		0.38	0.40	0.12	1.356	1.132	0.271
43	MC		0.40	0.27	0.22	1.349	0.985	0.257
44	MC	R	0.31	0.12	0.25	0.825	1.253	0.170
45	MC	R	0.61	0.13	0.16	0.692	0.755	0.176
46	MC		0.50	0.60	0.20	1.808	0.630	0.260
47	MC		0.62	0.38	0.16	1.082	0.469	0.307
48	MC		0.45	0.49	0.26	1.154	0.811	0.275
49	MC		0.51	0.51	0.23	1.487	0.696	0.315
50	MC	D	0.27	0.33	0.46	1.130	1.959	0.265

Appendix B: Flagged Item Summaries

The tables in Appendix B present the number of items flagged based on the following criteria:

- A = p-value < 0.20;
- R = discrimination < 0.25;
- D = distractor chosen by > 20% of high ability test takers; and
- H = p-value > 0.90.

The flag for omitted items is not included in these tables because no items were flagged due to a high omit rate (5% for the multiple choice items and 15% for the essays).

Table B.1 Flagged MC Items, by Form: Reading

	Total Number of Items	A Flag		R Flag		D Flag		H Flag	
		N	%	*N*	%	*N*	%	*N*	%
Form A	40	0	0.0	3	7.5	0	0.0	4	10.0
Form B	40	0	0.0	1	2.5	0	0.0	4	10.0
Form C	40	0	0.0			0	0.0	2	5.0

Table B.2 Flagged MC Items, by Form: Writing

	Total Number of Items	A Flag		R Flag		D Flag		H Flag	
		N	%	*N*	%	*N*	%	*N*	%
Form A	50	0	0.0	7	14.0	1	2.0	4	8.0
Form B	50	0	0.0	1	2.0	0	0.0	3	6.0
Form C	50	0	0.0	3	6.0	1	2.0	3	6.0

Table B.3 Flagged MC Items, by Form: Mathematics

	Total Number of Items	A Flag		R Flag		D Flag		H Flag	
		N	%	*N*	%	*N*	%	*N*	%
Form A	50	12	24.0	11	22.0	9	18.0	0	0.0
Form B	50	12	24.0	15	30.0	14	28.0	0	0.0
Form C	50	17	34.0	11	22.0	16	32.0	0	0.0

Table B.4 Flagged MC Items, by Form: Science									
	Total Number of Items	A Flag		R Flag		D Flag		H Flag	
		N	%	*N*	%	*N*	%	*N*	%
Form A	50	1	2.0	5	10.0	2	4.0	5	10.0
Form B	50	2	4.0	5	10.0	2	4.0	0	0.0
Form C	50	0	0.0	5	10.0	0	0.0	2	4.0

Table B.5 Flagged MC Items, by Form: Social Studies									
	Total Number of Items	A Flag		R Flag		D Flag		H Flag	
		N	%	*N*	%	*N*	%	*N*	%
Form A	50	0	0.0	4	8.0	2	4.0	3	6.0
Form B	50	1	2.0	4	8.0	0	0.0	1	2.0
Form C	50	0	0.0	6	12.0	1	2.0	2	4.0

Appendix C: Summary Item Statistics, By Form

Table C.1 Summary of Multiple-choice Item Statistics, by Form: Reading

	p-value	Discrimination	Parameter Estimates		
			a	*b*	*c*
Form A					
Number of items	40	40	40	40	40
Mean	0.70	0.49	1.20	0.04	0.22
Median	0.71	0.51	1.16	0.12	0.23
Standard deviation	0.15	0.11	0.43	0.67	0.06
Minimum	0.34	0.16	0.47	-1.61	0.06
Maximum	0.93	0.66	2.06	1.17	0.32
Form B					
Number of items	40	40	40	40	40
Mean	0.70	0.47	1.02	0.02	0.23
Median	0.74	0.49	0.99	-0.05	0.22
Standard deviation	0.16	0.10	0.33	0.69	0.09
Minimum	0.28	0.19	0.29	-1.27	0.01
Maximum	0.94	0.64	1.61	2.00	0.50
Form C					
Number of items	40	40	40	40	40
Mean	0.67	0.52	1.37	0.02	0.25
Median	0.68	0.53	1.11	0.20	0.25
Standard deviation	0.16	0.06	0.69	0.81	0.05
Minimum	0.34	0.38	0.56	-1.89	0.15
Maximum	0.96	0.65	3.36	1.21	0.40

Table C.2 Summary of Multiple-choice Item Statistics, by Form: Writing					
			Parameter Estimates		
	p-value	Discrimination	*a*	*b*	*c*
Form A					
Number of items	50	50	50	50	50
Mean	0.63	0.40	1.14	0.18	0.22
Median	0.67	0.42	1.12	0.15	0.22
Standard deviation	0.19	0.12	0.39	0.73	0.07
Minimum	0.23	0.02	0.39	-1.44	0.08
Maximum	0.94	0.58	1.89	1.93	0.37
Form B					
Number of items	50	50	50	50	50
Mean	0.64	0.40	1.15	0.19	0.24
Median	0.63	0.41	1.09	0.24	0.24
Standard deviation	0.17	0.09	0.47	0.76	0.07
Minimum	0.22	0.18	0.53	-1.43	0.05
Maximum	0.92	0.58	3.53	1.92	0.39
Form C					
Number of items	50	50	50	50	50
Mean	0.63	0.42	1.18	0.21	0.23
Median	0.59	0.42	1.1	0.19	0.24
Standard deviation	0.17	0.10	0.44	0.60	0.06
Minimum	0.29	0.13	0.44	-1.22	0.04
Maximum	0.94	0.62	2.24	2.13	0.35

Table C.3 Summary of Multiple-choice Item Statistics, by Form: Mathematics					
			Parameter Estimates		
	p-value	Discrimination	*a*	*b*	*c*
Form A					
Number of items	50	50	50	50	50
Mean	0.36	0.37	1.13	0.94	0.16
Median	0.27	0.35	1.05	1.00	0.15
Standard deviation	0.21	0.14	0.39	0.87	0.06
Minimum	0.07	0.10	0.62	-0.89	0.03
Maximum	0.89	0.66	2.50	2.56	0.38
Form B					
Number of items	50	50	50	50	50
Mean	0.30	0.37	1.08	1.09	0.16
Median	0.25	0.35	1.02	1.13	0.16
Standard deviation	0.18	0.15	0.37	0.83	0.04
Minimum	0.04	0.09	0.50	-1.67	0.07
Maximum	0.85	0.65	2.32	2.70	0.25
Form C					
Number of items	50	50	50	50	50
Mean	0.32	0.36	1.16	1.06	0.16
Median	0.26	0.35	1.13	1.23	0.15
Standard deviation	0.21	0.13	0.34	0.93	0.07
Minimum	0.05	0.12	0.59	-1.59	0.03
Maximum	0.90	0.65	2.37	2.84	0.32

Table C.4 Summary of Multiple-choice Item Statistics, by Form: Science					
			Parameter Estimates		
	p-value	Discrimination	*a*	*b*	*c*
Form A					
Number of items	50	50	50	50	50
Mean	0.60	0.45	1.30	0.63	0.22
Median	0.57	0.48	1.20	0.68	0.22
Standard deviation	0.19	0.13	0.43	0.72	0.05
Minimum	0.19	0.13	0.73	-1.45	0.09
Maximum	0.96	0.71	2.25	2.17	0.33
Form B					
Number of items	50	50	50	50	50
Mean	0.58	0.45	1.35	0.76	0.22
Median	0.60	0.48	1.22	0.78	0.22
Standard deviation	0.19	0.13	0.55	0.58	0.07
Minimum	0.18	0.13	0.30	-0.40	0.01
Maximum	0.84	0.69	2.81	2.46	0.34
Form C					
Number of items	50	50	50	50	50
Mean	0.62	0.46	1.25	0.67	0.23
Median	0.61	0.49	1.20	0.80	0.22
Standard deviation	0.17	0.14	0.50	0.65	0.07
Minimum	0.29	0.01	0.28	-0.89	0.03
Maximum	0.92	0.67	2.55	2.07	0.38

Table C.5 Summary of Multiple-choice Item Statistics, by Form: Social Studies					
			Parameter Estimates		
	p-value	Discrimination	*a*	*b*	*c*
Form A					
Number of items	50	50	50	50	50
Mean	0.59	0.41	1.14	0.32	0.25
Median	0.58	0.43	1.07	0.47	0.26
Standard deviation	0.17	0.14	0.46	0.81	0.06
Minimum	0.23	-0.04	0.26	-1.31	0.06
Maximum	0.99	0.64	2.43	1.96	0.39
Form B					
Number of items	50	50	50	50	50
Mean	0.61	0.44	1.11	0.26	0.21
Median	0.62	0.46	1.08	0.36	0.22
Standard deviation	0.16	0.11	0.37	0.63	0.09
Minimum	0.16	0.20	0.39	-1.32	0.02
Maximum	0.94	0.64	2.46	1.52	0.41
Form C					
Number of items	50	50	50	50	50
Mean	0.57	0.42	1.16	0.45	0.24
Median	0.57	0.45	1.14	0.59	0.26
Standard deviation	0.16	0.13	0.37	0.65	0.08
Minimum	0.27	0.12	0.39	-1.32	0.03
Maximum	0.96	0.63	1.91	1.96	0.40

Appendix D: Test Taker Performance: English Paper, Spanish Online, and Spanish Paper

Table D.1 Total Test Scale Score Summary Statistics for Reading, by Demographic Group: English, Paper Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		28,225		11.51	12	3.99	1	20
Gender	Male	16,805	60	11.64	12	4.00	1	20
	Female	11,420	40	11.32	11	3.97	1	20
Race/ Ethnicity	American Indian	764	3	10.52	11	3.88	1	20
	Asian	540	2	9.62	10	4.38	1	20
	African American	4,612	16	9.75	10	3.74	1	20
	White	10,233	36	12.83	13	3.77	1	20
	Hispanic	6,018	21	10.75	11	3.78	1	20
	Pacific Islander	227	1	10.28	10	4.20	1	20
	Multiracial	685	2	12.69	13	3.82	2	20
	Other/ No Response	5,146	18	11.57	12	3.90	1	20

Table D.2 Total Test Scale Score Summary Statistics for Writing, by Demographic Group: English, Paper Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		25,978		11.89	12	2.95	2	20
Gender	Male	15,557	60	11.75	12	2.96	2	20
	Female	10,421	40	12.10	12	2.92	2	20
Race/ Ethnicity	American Indian	673	3	10.89	11	2.95	2	18
	Asian	467	2	11.15	11	3.46	2	20
	African American	3,981	15	10.94	11	2.76	3	20
	White	9,768	38	12.53	13	2.93	2	20
	Hispanic	5,487	21	11.47	11	2.83	2	20
	Pacific Islander	219	1	11.47	12	3.18	3	19
	Multiracial	664	3	12.72	13	2.91	4	20
	Other/ No Response	4,719	18	11.96	12	2.86	2	20

Table D.3 Total Test Scale Score Summary Statistics for Mathematics, by Demographic Group: English, Paper Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		30,631		8.87	9	3.69	1	20
Gender	Male	17,924	59	9.25	9	3.72	1	20
	Female	12,707	41	8.34	8	3.57	1	20
Race/ Ethnicity	American Indian	757	2	8.37	8	3.72	1	20
	Asian	512	2	10.23	10	4.62	1	20
	African American	5,257	17	7.56	7	3.31	1	20
	White	11,060	36	9.62	9	3.76	1	20
	Hispanic	6,428	21	8.44	8	3.50	1	20
	Pacific Islander	238	1	7.85	7	3.49	1	18
	Multiracial	771	3	9.56	9	3.90	1	20
	Other/ No Response	5,608	18	9.01	9	3.51	1	20

Table D.4 Total Test Scale Score Summary Statistics for Science, by Demographic Group: English, Paper Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		26,443		12.24	12	3.86	1	20
Gender	Male	15,791	60	12.68	13	3.88	1	20
	Female	10,652	40	11.58	12	3.74	1	20
Race/ Ethnicity	American Indian	711	3	11.36	11	3.78	2	20
	Asian	448	2	11.75	12	3.94	2	20
	African American	4,206	16	10.29	10	3.45	1	20
	White	9,753	37	13.54	14	3.70	1	20
	Hispanic	5,549	21	11.44	11	3.62	1	20
	Pacific Islander	232	1	11.25	11	3.47	3	20
	Multiracial	679	3	13.32	14	3.90	1	20
	Other/ No Response	4,865	18	12.29	13	3.78	1	20

Table D.5 Total Test Scale Score Summary Statistics for Social Studies, by Demographic Group: English, Paper Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		28,421		11.15	11	4.20	1	20
Gender	Male	16,689	59	11.75	12	4.24	1	20
	Female	11,732	41	10.29	10	3.98	1	20
Race/ Ethnicity	American Indian	767	3	10.12	10	3.94	2	20
	Asian	505	2	10.30	10	4.21	2	20
	African American	4,656	16	9.21	9	3.59	1	20
	White	10,287	36	12.41	13	4.19	1	20
	Hispanic	6,059	21	10.47	10	3.95	1	20
	Pacific Islander	259	1	9.68	10	3.98	1	20
	Multiracial	707	2	12.25	12	4.18	3	20
	Other/ No Response	5,181	18	11.35	11	4.17	1	20

Table D.6 Percentage of English, Paper Test Takers in each Performance Level: Reading					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	28,225		17	61	22
Gender					
Male	16,805	60	17	61	23
Female	11,420	40	18	62	20
Race/Ethnicity					
American Indian	764	3	24	62	14
Asian	540	2	35	52	13
African American	4,612	16	29	62	9
White	10,233	36	9	58	33
Hispanic	6,018	21	21	65	15
Pacific Islander	227	1	27	58	15
Multiracial	685	2	12	58	31
Other/No Response	5,146	18	16	63	21

Table D.7 Percentage of English, Paper Test Takers in each Performance Level: Writing					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	25,978		8	74	17
Gender					
Male	15,557	60	9	75	16
Female	10,421	40	7	74	19
Race/Ethnicity					
American Indian	673	3	17	75	8
Asian	467	2	17	68	15
African American	3,981	15	12	79	9
White	9,768	38	6	71	23
Hispanic	5,487	21	9	77	13
Pacific Islander	219	1	13	73	14
Multiracial	664	3	5	69	26
Other/No Response	4,719	18	8	75	17

Table D.8 Percentage of English, Paper Test Takers in each Performance Level: Mathematics					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	30,631		39	54	7
Gender					
Male	17,924	59	35	57	9
Female	12,707	41	44	51	5
Race/Ethnicity					
American Indian	757	2	45	48	7
Asian	512	2	30	50	20
African American	5,257	17	53	44	3
White	11,060	36	31	59	10
Hispanic	6,428	21	42	52	5
Pacific Islander	238	1	50	46	3
Multiracial	771	3	32	57	11
Other/No Response	5,608	18	36	58	6

Table D.9 Percentage of English, Paper Test Takers in each Performance Level: Science					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	26,443		11	59	29
Gender					
Male	15,791	60	10	57	34
Female	10,652	40	14	64	23
Race/Ethnicity					
American Indian	711	3	16	63	22
Asian	448	2	14	61	25
African American	4,206	16	20	69	11
White	9,753	37	6	51	43
Hispanic	5,549	21	13	66	21
Pacific Islander	232	1	13	69	18
Multiracial	679	3	8	51	41
Other/No Response	4,865	18	10	61	29

Table D.10 Percentage of English, Paper Test Takers in each Performance Level: Social Studies

	N	Percent of Total	Performance Levels		
			Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	28,421		22	55	24
Gender					
Male	16,689	59	18	54	29
Female	11,732	41	28	56	17
Race/Ethnicity					
American Indian	767	3	28	55	17
Asian	505	2	29	53	18
African American	4,656	16	36	55	9
White	10,287	36	14	52	34
Hispanic	6,059	21	25	57	18
Pacific Islander	259	1	33	55	12
Multiracial	707	2	16	51	33
Other/No Response	5,181	18	20	56	24

Table D.11 Total Test Scale Score Summary Statistics for Reading, by Demographic Group: Spanish, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		1,295		9.04	9	3.56	1	20
Gender	Male	463	36	9.14	9	3.65	1	19
	Female	832	64	8.98	9	3.51	1	20
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	2	> 1	*	*	*	*	*
	White	6	> 1	*	*	*	*	*
	Hispanic	1,256	97	9.04	9	3.57	1	20
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	0	-	-	-	-	-	-
	Other/ No Response	31	2	8.45	8	3.02	2	15

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.12 Total Test Scale Score Summary Statistics for Writing, by Demographic Group: Spanish, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		971		10.45	10	2.96	1	18
Gender	Male	356	37	10.15	10	3.03	3	18
	Female	615	63	10.62	11	2.91	1	18
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	2	> 1	*	*	*	*	*
	White	4	> 1	*	*	*	*	*
	Hispanic	939	97	10.45	10	2.96	1	18
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	0	-	-	-	-	-	-
	Other/ No Response	26	3	9.62	9.5	2.71	4	15

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.13 Total Test Scale Score Summary Statistics for Mathematics, by Demographic Group: Spanish, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		1,322		7.41	7	3.46	1	20
Gender	Male	485	37	7.85	7	3.75	1	20
	Female	837	63	7.15	7	3.26	1	18
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	2	> 1	*	*	*	*	*
	White	7	> 1	*	*	*	*	*
	Hispanic	1,273	96	7.40	7	3.46	1	20
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	0	-	-	-	-	-	-
	Other/ No Response	40	3	7.05	7	3.11	2	13

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.14 Total Test Scale Score Summary Statistics for Science, by Demographic Group: Spanish, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		1,098		9.90	10	3.39	1	19
Gender	Male	390	36	10.55	10	3.62	2	19
	Female	708	64	9.54	9	3.21	1	19
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	2	> 1	*	*	*	*	*
	White	5	> 1	*	*	*	*	*
	Hispanic	1,063	97	9.90	10	3.4	1	19
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	0	-	-	-	-	-	-
	Other/ No Response	28	3	9.11	8.5	2.59	5	14

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.15 Total Test Scale Score Summary Statistics for Social Studies, by Demographic Group: Spanish, Online Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		1,225		8.95	9	3.40	2	20
Gender	Male	412	34	9.55	9	3.80	2	20
	Female	813	66	8.64	8	3.14	2	19
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	2	> 1	*	*	*	*	*
	White	5	> 1	*	*	*	*	*
	Hispanic	1,190	97	8.95	9	3.40	2	20
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	0	-	-	-	-	-	-
	Other/ No Response	28	2	8.14	8	2.86	3	13

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.16 Percentage of Spanish, Online Test Takers in each Performance Level: Reading

			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	1,295		36	59	5
Gender					
Male	463	36	35	60	5
Female	832	64	36	59	5
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	2	> 1	*	*	*
White	6	> 1	*	*	*
Hispanic	1,256	97	36	59	5
Pacific Islander	0	-	-	-	-
Multiracial	0	-	-	-	-
Other/No Response	31	2	35	61	3

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.17 Percentage of Spanish, Online Test Takers in each Performance Level: Writing

			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	971		21	70	9
Gender					
Male	356	37	24	68	7
Female	615	63	19	72	9
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	2	> 1	*	*	*
White	4	> 1	*	*	*
Hispanic	939	97	21	71	8
Pacific Islander	0	-	-	-	-
Multiracial	0	-	-	-	-
Other/No Response	26	3	38	58	4

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.18 Percentage of Spanish, Online Test Takers in each Performance Level: Mathematics					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	1,322		56	41	3
Gender					
Male	485	37	52	42	6
Female	837	63	58	40	2
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	2	> 1	*	*	*
White	7	> 1	*	*	*
Hispanic	1,273	96	56	41	4
Pacific Islander	0	-	-	-	-
Multiracial	0	-	-	-	-
Other/No Response	40	3	60	40	0

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.19 Percentage of Spanish, Online Test Takers in each Performance Level: Science					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	1,098		24	67	9
Gender					
Male	390	36	21	65	14
Female	708	64	25	69	6
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	2	> 1	*	*	*
White	5	> 1	*	*	*
Hispanic	1,063	97	24	67	9
Pacific Islander	0	-	-	-	-
Multiracial	0	-	-	-	-
Other/No Response	28	3	21	79	0

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.20 Percentage of Spanish, Online Test Takers in each Performance Level: Social Studies					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	1,225		36	57	7
Gender					
Male	412	34	32	56	12
Female	813	66	38	58	4
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	2	> 1	*	*	*
White	5	> 1	*	*	*
Hispanic	1,190	97	36	57	7
Pacific Islander	0	-	-	-	-
Multiracial	0	-	-	-	-
Other/No Response	28	2	36	64	0

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.21 Total Test Scale Score Summary Statistics for Reading, by Demographic Group: Spanish, Paper Test Takers								
		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		3,020		9.21	9	3.29	1	19
Gender	Male	945	31	9.33	9	3.40	1	19
	Female	2,075	69	9.15	9	3.24	1	19
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	1	> 1	*	*	*	*	*
	White	6	> 1	*	*	*	*	*
	Hispanic	2,839	94	9.21	9	3.30	1	19
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	1	> 1	*	*	*	*	*
	Other/ No Response	173	6	9.18	9	3.19	3	16

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.22 Total Test Scale Score Summary Statistics for Writing, by Demographic Group: Spanish, Paper Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		2,576		10.74	11	2.79	1	20
Gender	Male	839	33	10.26	10	2.77	1	18
	Female	1,737	67	10.98	11	2.78	2	20
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	1	> 1	*	*	*	*	*
	White	5	> 1	*	*	*	*	*
	Hispanic	2,452	95	10.72	11	2.80	1	20
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	1	> 1	*	*	*	*	*
	Other/ No Response	117	5	11.18	11	2.71	5	16

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.23 Total Test Scale Score Summary Statistics for Mathematics, by Demographic Group: Spanish, Paper Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		3,101		8.05	8	3.41	1	20
Gender	Male	921	30	8.78	9	3.49	1	19
	Female	2,180	70	7.74	7	3.33	1	20
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	0	-	-	-	-	-	-
	White	8	> 1	*	*	*	*	*
	Hispanic	2,947	95	8.05	8	3.42	1	20
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	1	> 1	*	*	*	*	*
	Other/ No Response	145	5	8.06	8	3.25	1	18

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.24 Total Test Scale Score Summary Statistics for Science, by Demographic Group: Spanish, Paper Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		2,668		10.12	10	3.37	2	20
Gender	Male	818	31	10.83	11	3.48	2	19
	Female	1,850	69	9.81	10	3.27	2	20
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	0	-	-	-	-	-	-
	White	5	> 1	*	*	*	*	*
	Hispanic	2,517	94	10.11	10	3.38	2	20
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	1	> 1	*	*	*	*	*
	Other/ No Response	145	5	10.33	10	3.10	2	18

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.25 Total Test Scale Score Summary Statistics for Social Studies, by Demographic Group: Spanish, Paper Test Takers

		N	Percent of Total	Mean	Median	*SD*	Obs. Min.	Obs. Max.
Total		2,899		9.30	9	3.34	1	20
Gender	Male	884	30	9.97	10	3.52	1	20
	Female	2,015	70	9.01	9	3.22	1	20
Race/ Ethnicity	American Indian	0	-	-	-	-	-	-
	Asian	0	-	-	-	-	-	-
	African American	0	-	-	-	-	-	-
	White	8	> 1	*	*	*	*	*
	Hispanic	2,740	95	9.31	9	3.35	1	20
	Pacific Islander	0	-	-	-	-	-	-
	Multiracial	1	> 1	*	*	*	*	*
	Other/ No Response	150	5	9.20	9	3.19	3	17

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.26 Percentage of Spanish, Paper Test Takers in each Performance Level: Reading

			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	3,020		32	63	5
Gender					
Male	945	31	32	62	6
Female	2,075	69	32	63	4
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	1	> 1	*	*	*
White	6	> 1	*	*	*
Hispanic	2,839	94	32	63	5
Pacific Islander	0	-	-	-	-
Multiracial	1	> 1	*	*	*
Other/No Response	173	6	33	64	3

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.27 Percentage of Spanish, Paper Test Takers in each Performance Level: Writing

			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	2,576		15	77	8
Gender					
Male	839	33	20	75	5
Female	1,737	67	12	78	9
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	1	> 1	*	*	*
White	5	> 1	*	*	*
Hispanic	2,452	95	15	77	8
Pacific Islander	0	-	-	-	-
Multiracial	1	> 1	*	*	*
Other/No Response	117	5	9	81	9

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.28 Percentage of Spanish, Paper Test Takers in each Performance Level: Mathematics					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	3,101		48	48	4
Gender					
Male	921	30	40	54	6
Female	2,180	70	51	46	3
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	0	-	-	-	-
White	8	> 1	*	*	*
Hispanic	2,947	95	48	48	4
Pacific Islander	0	-	-	-	-
Multiracial	1	> 1	*	*	*
Other/No Response	145	5	44	52	4

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.29 Percentage of Spanish, Paper Test Takers in each Performance Level: Science					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	2,668		22	67	11
Gender					
Male	818	31	18	67	16
Female	1,850	69	24	67	9
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	0	-	-	-	-
White	5	> 1	*	*	*
Hispanic	2,517	94	23	66	11
Pacific Islander	0	-	-	-	-
Multiracial	1	> 1	*	*	*
Other/No Response	145	5	18	73	9

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

Table D.30 Percentage of Spanish, Paper Test Takers in each Performance Level: Social Studies					
			Performance Levels		
	N	Percent of Total	Did Not Pass (%)	Pass But Not CCR (%)	College Career Ready (%)
Total	2,899		33	59	8
Gender					
Male	884	30	27	62	12
Female	2,015	70	36	58	6
Race/Ethnicity					
American Indian	0	-	-	-	-
Asian	0	-	-	-	-
African American	0	-	-	-	-
White	8	> 1	*	*	*
Hispanic	2,740	95	33	59	8
Pacific Islander	0	-	-	-	-
Multiracial	1	> 1	*	*	*
Other/No Response	150	5	36	59	5

Note. Statistics not reported for sample size less than 25 (N < 25), denoted by '*'.

For more information,
Visit: **hiset.org**
Phone Toll-Free: **1-855-MyHiSET**
(1-855-694-4738)

801453