Technical Manual

2014 GED® Test

Table of Contents

Acknowledgments

GED Testing Service is pleased to issue the GED® Test Technical Manual. We wish to acknowledge the work of all those who contributed to this manual.

GED Testing Service

Martin Kehe, *Vice President, Assessment Services*

Tracy Gardner, *Senior Director, Assessment Services*

J. Carl Setzer, *Senior Psychometrician*

Yung-chen Hsu, *Senior Psychometrician*

Marjorie Wine, *Principal Test Development Manager*

Michael Bell, *Senior Content Specialist—Mathematics*

Mariana Holliday, *Content Specialist—Science*

James McKinney, *Senior Content Specialist—English Language Arts*

Bridget Olson, *Content Specialist—Social Studies*

Pearson School

Anna Topczewski, *Associate Research Scientist*

Pearson VUE

Linda Conheady, *Test and Report Editor*

Technical Advisory Committee

Neal Kingston, *Director, Achievement and Assessment Institute*

Ric Luecht, *Professor of Educational Research Methodology at University of North Carolina at Greensboro*

Bill Mehrens, *Professor Emeritus at Michigan State University*

Barbara Plake, *Professor Emeritus at University of Nebraska–Lincoln*

Cyndie Schmeiser, *Consultant and Chief of Assessment at College Board*

Preface

This manual was written to provide technical information regarding the GED® test as evidence that the GED® test is technically sound. Throughout this manual, documentation is provided regarding the development of the GED® test and data collection activities, as well as evidence of reliability and validity.

This manual is made up of 10 chapters, which include the following information:

Chapter 1: Introduction to the GED® tests and an overview of the GED® testing program, including the purposes of the tests and proper uses of test scores

Chapter 2: The underlying theory of action for the GED® testing program and the framework for assessing validity

Chapter 3: The GED® test specifications and process for developing forms

Chapter 4: The standardization process, including the norming, scaling, and equating processes

Chapter 5: The standard-setting process for both the High School Passing Standard and GED® with Honors

Chapter 6: Scoring processes for both machine- and human-scored test items

Chapter 7: The reliability of GED® test scores

Chapter 8: Accumulated evidence to support the validity argument

Chapter 9: Various accommodations for test-takers with disabilities

Chapter 10: Supplemental materials for test preparation

This manual was written for anyone who is interested in (a) learning about the background of the GED® testing program, (b) understanding how the GED® test was developed and scored, (c) comprehending the statistical characteristics of the GED® test, or (d) knowing more, in general, about the GED® testing program. Individuals interested in additional information are encouraged to visit the GED Testing Service web site at www.GEDtestingservice.com.

Chapter 1: Introduction

About the GED Testing Program

The GED Testing Service is a joint venture of the American Council on Education (ACE) and Pearson. Our mission, vision, and values are based on inclusiveness and diversity. We recognize the responsibility of those in the educational community to contribute to society. We also embrace the belief that widespread access to post-secondary education, particularly for those adult learners who seek lifelong learning, is the cornerstone of a democratic society.

GED Testing Service Vision

In an ideal society, everyone would graduate from high school. Until that becomes a reality, the GED Testing Service will offer the opportunity to earn a high school equivalency credential so that individuals can have a second chance to advance their educational, personal, and professional aspirations.

GED Testing Service Mission

GED Testing Service stands as the only time-honored architect of the GED® test, which certifies the high school–level academic achievement of national and international non–high school graduates. In collaboration with key partners, we develop, deliver, and safeguard our tests; we analyze the testing program and its participants; and we develop policies, procedures, and programs to ensure equal access to our tests.

GED Testing Service Values

The integrity of GED Testing Service and its products rests on our commitment to excellence, diversity, inclusiveness, educational opportunities, and lifelong learning. This commitment is reflected in our proactive approach to developing collaborative solutions, our research-based decision making, and our timely support to the people we serve.

History of the GED® Tests

The first GED® tests were developed in 1942 to measure the major outcomes and concepts generally associated with four years of high school education. Initiated by the United States Armed Forces Institute (USAFI), the original tests were administered only to military personnel so that returning World War II veterans could more easily pursue their educational, vocational, and personal goals.

The USAFI examination staff, composed of civilian testing experts, worked with an advisory committee established with the support and cooperation of ACE, the National Association of Secondary School Principals, and regional U.S. accrediting associations. Lindquist (1944) paved the way for the GED Tests by establishing a philosophical and technical basis for exam-based equivalency. In 1945, ACE established the Veterans' Testing Service (VTS; predecessor of today's GED Testing Service). The VTS took over the development and administration of the GED® tests and focused on helping World War II veterans pursue educational and vocational goals without returning them to the classroom.

The opportunity to document the attainment of high school–level academic skills served as a significant aid to the many service members whose academic careers had been disrupted during the war. During the 1950s, it became apparent that civilians could also benefit from the program—a need that ACE undertook to fulfill. New York was the first state to allow nonveteran adults to take the GED® tests, in late 1947. In 1955, policies regarding administration of the tests in federal correctional and health institutions were modified. By 1959, more nonveterans than veterans were taking the GED® tests. With the growth of the high school equivalency program, ACE made the decision in 1960 to transfer the college-level GED® tests to the Educational Testing Service (ETS). Those tests that ETS developed are known today as part of the College Board's College-Level Examination Program® (CLEP).

From 1945 to 1963, the program was administered by the VTS. In 1958, a policy to allow overseas testing of U.S. civilians and foreign nationals was approved. In 1963, in recognition of the transition to a program chiefly for nonveteran adults, the name was changed to GED Testing Service. To serve all qualified examinees equally, the Commission on Accreditation of Service Experiences approved English-language versions in audio, Braille, and large-print formats in 1964. In addition, Nova Scotia became the first Canadian province to offer GED® testing to civilians in 1969, and in 1970, the first English-language Canadian version of the GED® tests was published. In 1973, GED Testing Service reached a milestone when California became the last state to adopt a uniform acceptance of the GED® tests. (For more history on the GED® tests, see Mullane [2001].)

The five original GED® tests in use from 1942 to 1978 were titled:

Test 1: Correctness and Effectiveness of Expression
Test 2: Interpretation of Reading Materials in the Social Studies
Test 3: Interpretation of Reading Materials in the Natural Sciences
Test 4: Interpretation of Literacy Materials
Test 5: General Mathematical Ability

The entire battery took 10 hours to administer. (For more information about the content of the first generation of GED® tests, see American Council on Education [1964].) In the 1970s it became apparent that the effects of changed secondary curricula and, perhaps, changed attitudes toward education among the general public necessitated a review of the specifications of the GED® tests. This review resulted in a thorough revision of the first series of GED® tests.

The second series of GED® tests, introduced in 1978, was based on test specifications defined in the mid-1970s by committees of high school curriculum specialists. Among the major changes was the development of a Reading Skills test to replace Test 4: Interpretation of Literary Materials, and the reduction of the reading load in the Science and Social Studies tests. In addition, “concept” items were developed to make up one-third of the Science and Social Studies tests. These items required much less reading than the reading comprehension items, which dominated previous tests, but it was assumed that examinees had some prior knowledge in science and social studies. In addition, Test 1: Correctness and Effectiveness of Expression, was replaced by a Writing Skills test. The Mathematics test included more practically-oriented test items.

The second series of GED® tests originally required 6 hours of test administration time. On the basis of research (Whitney & Patience, 1981), the Commission on Accreditation of Service Experiences in 1981 increased the time limits for the Writing Skills Test from 60 minutes to 75 minutes and for the Mathematics Test from 60 minutes to 90 minutes.

This series of tests, used from 1979 through 1987, consisted of the following titles:

Test 1: The Writing Skills Test
Test 2: The Social Studies Test
Test 3: The Science Test
Test 4: The Reading Skills Test
Test 5: The Mathematics Test

These tests retained the emphasis on demonstrating the designated high school outcomes but introduced “real-life” contexts into many of the test items. They also introduced many reading materials likely to be encountered in an adult’s daily life (such as schedules and newspaper articles). (For further information about the development of the second-generation tests, including detailed descriptions, see Patience and Whitney [1982].)

The development of the third series of GED® tests, used from 1988 through 2001, began in November 1982 in order to ensure that the GED® tests addressed and measured the educational academic outcomes expected of graduating high school seniors during the late 1980s and early 1990s.

The Tests Specifications Committee offered recommendations for the entire GED® test battery centered on five major themes:

- Requiring examinees to demonstrate their higher-level thinking abilities and problem-solving skills
- Including a clear emphasis on the relationship of the skills tested to aspects of the world of work
- Representing an awareness of the role and impact of computer technology
- Addressing certain consumer skills, and
- Using contexts that adult examinees would recognize and include stimulus materials related to aspects of everyday life.

The GED® tests introduced in 1988:

- Required a direct writing sample.
- Demanded more highly developed levels of critical thinking
- Reflected many roles of adults
- Acknowledged the sources of change affecting individuals and society
- Contained contexts that adult examinees would recognize as relevant to daily life

This third series of GED® tests required 7 hours and 45 minutes of test administration time. The official titles of the five separate subject tests were:

Test 1: Writing Skills
Test 2: Social Studies
Test 3: Science
Test 4: Interpreting Literature and the Arts
Test 5: Mathematics

In the late 1990s, GED Testing Service undertook a study comparing national and state standards in English language arts, mathematics, science, and the disciplines within social studies to survey what graduating students should know and be able to do (American Council on Education, 1999). The major purpose of the study was to inform the education community and the public of the development of the fourth series of GED® tests. By identifying the common elements among national and state standards and aligning the test specifications to these standards, GED Testing Service provided support for the claim that the GED® test scores can be used to measure what adult examinees know and are able to do relative to high school seniors who are completing a traditional four-year program of study.

This fourth series of GED® Tests (English-language U.S. and Canadian versions) was released in 2002.[1] The official titles of the five separate subject tests were:

Language Arts, Writing
Social Studies
Science
Language Arts, Reading
Mathematics

In 2008, GED Testing Service once again began the process of developing a new test.[2] Around that same time, the educational climate had begun a shift toward academic standards that were more closely aligned with career- and college-readiness standards. GED Testing Service worked with a number of external consultants to determine its strategic course of action regarding the direction of the program. Ultimately, it was decided that adults without a high school diploma need to have the same opportunities to demonstrate their readiness for careers and post-secondary education.

[1] In addition to the standard print editions, large-print, audiocassette, and Braille editions were introduced for the 2002 Series at the same time. These editions were developed for those in need of special accommodations.

[2] Historically, GED Testing Service has referred to the GED® Tests (plural). With the launch of the new series in 2014, the new test is referred to in singular form as the GED® test.

Overview of the GED® Test

GED Testing Program Assessments

The GED® testing program comprises two assessments: GED Ready®: The Official Practice Test and the flagship GED® test. These tests were developed concurrently, with similar specifications, yet are utilized differently. GED Ready® was built to the same content specifications as the GED® test and therefore allows individuals to evaluate their likelihood of success on the GED® test. GED Ready® also provides an opportunity to gain additional testing experience as well as exposure to the content and format of the GED® test.

The GED® test is intended for jurisdictions to award a high school–level credential. The GED® test is also an indication of readiness for some jobs, job training programs, and/or credit-bearing post-secondary education coursework.

Both the GED® test and GED Ready® are primarily administered via computer.[3] Each test is promptly scored by a secure system so that results may be reported to the candidate and to other relevant agencies or organizations in a timely manner. Writing prompts and performance tasks are scored using computer-based, automated scoring engines to facilitate prompt reporting.

Both tests utilize a variety of item formats to tap into a range and depth of knowledge and skills. The use of multiple item types is intended to improve measurement of test-takers' true abilities, to make the test more engaging and more reflective of real-world tasks and situations, and to increase test-taker motivation. Both tests also leverage technology for test administration, scoring, and reporting purposes. By administering computer-based versions of each test, GED Testing Service can standardize the test administration process, score items, and provide interpretable feedback to the test-takers and various stakeholder groups who use the results.

GED® Test Purpose

The philosophy underlying the GED® test is that there is a foundational core, or *domain*, of academic skills and content knowledge that must be demonstrated for an adult to be certified as prepared to enter a job, a training program, or an entry-level, credit-bearing post-secondary course. This foundational core of knowledge and skills is defined by career- and college-readiness standards, now adopted in some form by the majority of states.

At the same time, GED Testing Service recognizes that high school graduation requirements vary from state to state, and perhaps even district to district within states. Additionally, there are varying degrees of career- and college-readiness. GED Testing Service must ensure that we hold adults in need of a high school credential to similar expectations of traditional high school students while also recognizing variability in performance.

[3] Paper, Braille, screen reader and other accommodated versions of the test are also available in certain circumstances.

The GED® test has traditionally been, and will continue to be, the flagship product of the GED Testing Service. The GED® test provides a second opportunity for millions of adults who did not receive a traditional high school diploma to obtain a high school–level credential. In previous GED® test series, the foremost purpose of the GED® test was to provide jurisdictions a mechanism in which to award their high school–level credential. GED Testing Service seeks to continue and strengthen the quality of this product.

The GED® test has three purposes. These purposes and their intended uses are listed below:

1. **To provide a high school–level credential.** The GED® test provides an alternative pathway for adults who have not earned a high school diploma by measuring whether a test-taker has acquired the foundational academic skills and knowledge necessary for high school exit at a degree of performance consistent with that demonstrated by graduating high school seniors. The intended use of the GED® test credential is similar to that of a high school diploma—to qualify for jobs and job promotions, to enable further education and training, and to enhance an adult's personal satisfaction.

2. **To provide information about a candidate's strengths and areas of developmental need.** By providing more information about candidate performance on career- and college-readiness standards, learning network counselors, post-secondary advisors, trainers, and educators can help candidates strengthen those areas of developmental need so that the candidates are ready for the next level of lifelong education and training.

3. **To provide evidence of readiness to enter workforce training programs, some careers, or post-secondary education.** The GED® test results indicate whether a candidate is likely ready to enter a workforce training program or a typical entry-level, credit bearing, 2- or 4-year college course without needing remediation.[4]

GED® test scores should not be used to make inferences regarding any non-cognitive aspects often developed by attending high school, such as creativity, teamwork, planning and organization, ethics, leadership, self-discipline, and socialization. In addition, GED Testing Service's policy clearly states that the GED® test should not be used to validate high school diplomas and does not permit the tests to be administered to high school students still enrolled in school or high school graduates, except under special circumstances. Employers and post-secondary institutions are explicitly forbidden to use the GED® test to verify the achievement level of high school graduates.

[4] A career- and college-readiness indicator is included in the enhanced score reporting. Currently, this performance level is called, "GED® with Honors." As further validity evidence is collected, this indicator is expected to serve as the basis of awarding a career- and college-readiness credential. Longitudinal data collected over time ultimately will be used to fulfill this purpose.

The GED® test is future focused: that is, it is designed to provide information about candidate readiness that is directly tied to the next steps in candidates' preparation and training. Many factors impact actual workplace or post-secondary success (such as engagement, teamwork, or creativity). However, the GED® test focuses on those foundational career and college academic skills and knowledge that are critical for candidates to be prepared for the next step in their future, whether they seek to enter the workforce or some form of post-secondary education for further education and training.

Intended Examinee Population

The GED® test is intended for adults who do not have a traditional high school diploma. The GED® test is not intended for high schools to use as an exit examination. However, there may be cases in which nontraditional students (e.g., home-schooled students) may take the GED® test as a means for obtaining a high school–level credential. Each jurisdiction maintains its own set of eligibility requirements and policies (see GED Testing Service, 2012).

Performance Benchmarks and Credentials

The GED® test offers two performance levels of passing scores. The first performance level determines whether candidates are eligible to receive their jurisdiction's high school–level credential. Those candidates who attain a score at or above the high school credential benchmark in *each* of the four content area tests become eligible to receive their jurisdiction's high school–level credential. The second performance level is used to determine whether the candidate has reached a level of performance indicative of career- and college-readiness. A candidate may receive a career- and college-readiness designation if that candidate (1) attains a score at or above the career- and college-readiness passing standard in that content area (known as "GED® with Honors"), and (2) has met the first benchmark (high school credential) in each of the four content areas.

The GED® Passing Standard

The GED® passing standard refers to the minimum level of performance in each content area that must be achieved in order for a GED® test-taker to be eligible to earn a GED® test credential.

The GED® passing standard is defined as *that point on the score scale that represents a reasonable and sufficient level of performance to expect from adult learners on the targeted knowledge and skills, given the performance of a national sample of graduating high school seniors*. In this context, "reasonable and sufficient" are defined as ***not too high*** as to hold adult learners to a higher standard than that of graduating high school seniors and ***not too low*** as to threaten the validity of the results for use in awarding a GED® test credential.

Given this definition of the passing standard, the inference associated with a test-taker who performs at or above the GED® passing standard is simply that he or she *achieved a sufficient level of performance within that content area to be eligible to earn a GED® test credential*.

The Career- and College-Readiness Indicator

The generation of a benchmark that provides test-takers with information regarding their "readiness" for careers and college is a component of the GED® test that is new to the current test edition. It is important that GED Testing Service be extremely clear about the meaning of this standard and the inferences it is intended to support at different points within the life cycle of the GED® test.

The career- and college-readiness (CCR) indicator is defined as *that point on the GED® test score scale that represents a level of performance estimated to be indicative of career- and college-readiness.*

Given this definition, the inference associated with performance at or above the CCR indicator is simply that a candidate attained a level of performance within this content area estimated to be indicative of career- and college-readiness.

As with the GED® passing standard, it is important to note the caveats implied by this definition. We are not stating that performance relative to the cut score provides for any predictive inferences regarding readiness for careers or college. Rather, we are claiming that meeting this performance benchmark indicates the test-taker has demonstrated a level of performance *consistent* with that suggested by other related measures. Collecting empirical evidence to fully support this claim will be done over time.

GED Testing Service acknowledges and understands that the operational definition of "readiness" is fluid over time. The national conversation regarding what constitutes readiness for careers and college will continue to evolve as research and policy continue to mount. In that sense, future changes to the CCR benchmark may be necessary.

Guiding Principles of the GED® Test Design

The four content area assessments (Reasoning Through Language Arts, Mathematical Reasoning, Science, Social Studies) measure a foundational core of knowledge and skills that are essential for career- and college-readiness. The domain assessed by each test is defined by a set of standards that describes what the candidate should know and be able to do. Each of the content area tests is briefly described below.

GED® Reasoning Through Language Arts

In alignment with career- and college-readiness standards, the GED® Reasoning Through Language Arts test focuses on three essential elements: (1) the ability to read closely, (2) the ability to write clearly, and (3) the ability to edit and understand the use of standard written English in context. Because the strongest predictor of career- and college-readiness is the ability to read and comprehend complex texts, especially in nonfiction, the Reasoning Through Language Arts test includes a range of texts from both academic and workplace contexts that vary from simple to more complex. The writing tasks require test-takers to analyze given source texts, using evidence drawn from the text(s).

GED® Mathematical Reasoning

The GED® Mathematical Reasoning test content focuses on those career- and college-readiness standards that are most important for high school exit and career- and college-readiness. The Mathematical Reasoning test focuses primarily on quantitative skills and algebraic problem solving and measures these skills in both academic and workplace contexts.

GED® Science

The GED® Science test measures the candidate's ability to demonstrate scientific literacy skills in the context of real-life science contexts. The skills are those that are important to career- and college-readiness and include quantitative reasoning within science and literacy skills, such as analysis of scientific and technical text sources that include data reported in graphs and charts. A focused selection of content (e.g., life science, physical science, and earth and space science) and key themes in science (health and the human body, energy) set the context within which these skills are measured.

GED® Social Studies

The GED® Social Studies test measures a candidate's ability to read, understand, analyze, and write about issues and topics in social studies. These skills are those that are important to career- and college-readiness and include literacy skills within social studies as well as numeracy and reasoning skills involving data, charts, graphs, and maps within social studies. Test-takers write a short analytic essay based on given source text(s). The primary content focuses on the United States founding documents and writings of major historical contributors, and to a lesser degree, on global topics.

Global Claims

Test scores require interpretation. Test-takers who take and pass all four content area tests need a clear understanding of the test score interpretations, as do stakeholders (e.g., employers and postsecondary education admissions offices). The following is a list of claims that can be used to interpret the performance of adults who have passed the GED® test at the battery level and are thus eligible to receive a high school credential.

Claim 1 Candidates can read closely and critically to comprehend a range of increasingly complex texts taken from literary, workplace, social studies, and science contexts.

Claim 2 Candidates can produce a clear, effective written analysis and reflection using evidence drawn from source texts provided.

Claim 3 Candidates can demonstrate command of the conventions of standard English when editing in authentic contexts and when producing written analysis.

Claim 4 Candidates can evaluate arguments, including the validity of reasoning and the relevance and sufficiency of evidence.

Claim 5 Candidates can explain and apply mathematical concepts and carry out mathematical procedures with precision and fluency. Contexts for which these skills and knowledge are applied include academic (traditional mathematics, social studies and science) sources as well as workplace contexts.

Claim 6 Candidates can frame, analyze, and solve a broad range of quantitative and algebraic problems in real-life, practical workplace, social studies, and science contexts.

Claim 7 Candidates can understand, interpret, analyze, summarize, and evaluate data provided in text and in various graphical forms pertaining to complex issues and problems in mathematics, social studies, and science.

Claim 8 Candidates can understand, apply, and critically analyze information and data related to the development, from ancient civilizations to the present, of current ideas about democracy and human and civil rights. Candidates can demonstrate how the systems, structures, and policies that people have created interact within four major content domains in social studies: civics and government, U.S. history, economics, and geography.

Claim 9 Candidates can understand, apply, and critically analyze scientific information and data related to the focusing themes of Health and Human Body and of Energy within three major content domains in science: life science, physical science, and earth and space science.

Item Formats

Various item formats are used throughout the content area tests. Each of the content area tests contains multiple-choice items, each containing four response options. Other item formats that may be included on various forms include fill-in-the-blank, drag-and-drop, hot spot, and/or drop-down items. The Reasoning Through Language Arts test also includes cloze items as well as an extended response item. The Social Studies test also includes an extended response item. The Science test includes short answer items. All items—including the extended response and short answer items—are scored electronically. More specific information on the item formats found on each content area test can be found in Chapter 3.

Accommodations

GED Testing Service has established procedures for adults with documented disabilities to obtain accommodations for the GED® test. GED Testing Service encourages individuals who may benefit from accommodations to take advantage of the opportunities available to them via their jurisdiction's GED® testing program. Accommodations are provided for adults with physical, learning, and psychological disabilities as well as those with attention-deficit/hyperactivity disorder. Approval for accommodations and use of special editions for adults with disabilities must be obtained through an accommodations request process.

Individuals with disabilities must be able to provide adequate documentation and must request accommodations through the GED Testing Service Web Portal. They are required to submit appropriate forms (based on the type of disability). Available accommodations for the computer-based GED® test include, but are not limited to, the following:

- Extended testing time (e.g., 25%, 50%, 100% additional time)
- Extra breaks
- A private testing room
- A paper version of the test for test takers that have a medical condition that precludes them from using a computer screen.
- An audio version of the test (provided through the computer using Job Access With Speech [JAWS] screen reading technology integrated into the computer test driver)
- Presentation of the test material on computer in an appropriate point size as determined by an individual test-taker using a “zoom” feature known as Zoom Text
- Braille
- Talking calculator for visually impaired test-takers
- Scribe and/or reader for test-takers with a variety of physical conditions that prohibit test-takers from reading or responding on their own

Test-takers also have two levels of font enlargement available to them. First, all test-takers have access to a functionality embedded within the test that allows them to enlarge the text to up to 20-point font. Second, test-takers who qualify for an additional accommodation can request access to ZoomText software, which allows them to enlarge fonts to virtually any size they desire. Both of these functionalities are available on all English and Spanish forms of the test.

Time Limits

The time limits for each of the test versions are provided in Table 1.

Table 1. Time Limits Applied to the GED® Test and GED Ready® (in minutes)

Content Area	GED® Test	GED Ready®
Reasoning Through Language Arts	150	95
Mathematical Reasoning	115	60
Science	90	47
Social Studies	90	60

The time limits shown in the table above refer to the total time allotted for each test. For each content area test, a fraction of the time is allotted for reading introductory and closing text. The RLA test comprises two sections with a ten-minute break in between. The first section contains a set of selected response items and the extended response item. The total allotted time for Part I is 72 minutes. After completing Part I, the test-taker is allowed to take a 10-minute break. Part II of the test is allotted 65 minutes. An additional 3 minutes on the RLA test is allotted to reviewing test instructions and the review and test submission screens. The Social Studies test also contains two parts. The first part contains selected response items only and is allotted 63 minutes. Part II of the Social Studies test contains the extended response item and is allotted 25 minutes. The Social Studies test provides for 2 minutes to be spent on the test instructions and the review and submission screens.

The GED Ready® versions of Reasoning Through Language Arts and Social Studies are structured differently. Part I of Reasoning Through Language Arts is allotted 47 minutes and contains selected response (multiple choice) and technology enhanced items. Part II is allotted 45 minutes for the extended response item. Part I of the Social Studies test is allotted 33 minutes and contains selected response and technology enhanced items; Part II is allotted 25 minutes for the extended response item.

Test Administration

The GED® test is available for administration in one of three types of sites: (a) Pearson VUE Authorized Testing Centers, (b) mobile sites, and (c) GED®-only sites. Pearson VUE testing centers (PVTCs) and their additional sites are fixed, physical testing sites independently authorized to offer a full range of Pearson VUE tests. A mobile site is a non-fixed, laptop-based testing platform that is linked to and dependent on another PVTC. Mobile sites are used to serve remote communities, correctional institution sites, and other locations that have less frequent testing needs. Test-takers wishing to utilize the mobile site can register on GED Testing Service's website but must make arrangements with the PVTC to utilize the mobile site. Finally, other fixed, physical testing sites may choose to only offer the GED® test.

At each site, a manual of guidelines for GED® test proctors and examiners describes how the GED® test is to be administered. This *Policies and Procedures* guide provides detailed instructions to test administrators regarding the use of the center's software applications and managing the test administration process. It is updated regularly and includes GED® test administration policies.

The vast majority of the test administration is by computer using Pearson VUE's Athena delivery platform. Accommodations are provided for paper, Braille, and other test versions as previously described.

Scoring Procedures

Most of the items on the GED® test are scored immediately and directly within the test delivery platform. The majority of the items are scored dichotomously (i.e., either correct or incorrect). However, in some cases there are polytomously scored items (i.e., partial credit is awarded). In other cases, items are scored dichotomously, yet are double-weighted (i.e., zero or two points are awarded). See Chapter 3 for additional details about items and point-value distributions.

Extended response (ER) and short answer (SA) items are externally scored using an automated scoring engine. Both the Reasoning Through Language Arts and the Social Studies ERs are analytically scored using a three-trait rubric and are double weighted. SA items on the Science test are polytomously scored on a scale of 0 to 3.

In the GED® test, the vast majority of test-taker responses to the ER and SA items (approximately 97 percent of the total responses) are scored within nanoseconds, and the score is returned to the reporting system. The remaining responses (approximately 3 percent) represent writing samples that are unique or have unusual characteristics that prevent them from being scored with a high degree of certainty by the automated scoring engine. These “outlier” responses are immediately flagged and sent via a secure encrypted process to expert human readers in a distributed scoring network at Pearson.

Two readers read each of the flagged outlier responses, and if those readers have exact agreement on the score or are within one (1) score point difference, the score is finalized at the higher of the two scores and returned to the reporting system. If the readers differ in their scores of a response by more than one (1) score point, the response goes to a third expert scorer, who adjudicates the response and finalizes the score for return to the reporting system. In addition, the ER and SA scoring is submitted to a 1 percent randomly sampled rescoring process in which human readers verify that the automated scoring engine is operating in accordance with the predetermined human scoring process that it was designed to replicate.

Explanation of Standard Scores

GED® test standard scores (i.e., scaled scores) were developed through a norming study (described in Chapter 4) involving a national sample of recent high school graduates. GED® standard scores thus provide a standard against which an adult’s test performance can be evaluated. This standard involves an external yardstick based on the achievement levels of contemporary high school seniors.

The process GED Testing Service uses to establish standard scores helps ensure that minor differences in difficulty among the various forms of the GED® tests will neither help nor hinder an examinee’s completion of a particular form. That is, standard scores are used to make appropriate adjustments for the fact that the items on some test forms may be slightly easier or slightly more difficult than those on another form (within the same content area). The use of standard scores ensures that an examinee can expect to earn about the same score regardless of test form.

The standard scores are used to compare the achievement of GED® test-takers directly with the demonstrated achievement of recent high school graduates. To qualify for a high school credential, a GED® test-taker must perform at least as well as a certain percentage of recently graduated high school seniors (see Performance Benchmarks and Credentials section).

In reporting scores earned on the GED® test, GED Testing Service uses standard scores and percentile ranks (the percentage of recent high school graduates who scored at or below that standard score). Both score scales involve transforming the test-taker’s raw score (number of items correctly answered) to new numerical scales. Higher raw scores are associated with higher standard scores and percentile ranks.

For each content area test, the standard scores range from 100 to 200. The high school credential passing standard is set at 150, and the GED® with Honors benchmark is set at 170.

Examinee Feedback

Each person who takes a GED® test receives an official score report. Generally, the score report is available within approximately 3 hours of completing the exam. Test-takers are notified via email that the score report is available and are directed to log into their account on the MyGED® portal at www.ged.com where the score report can be accessed.

The score report provides the standard score for the test, the percentile rank associated with the standard score, and feedback regarding the types of knowledge and skills typically seen at that standard score level. The test-taker is provided information on the skills he or she did well on, in addition to the specific skills the test-taker missed on the test. A list of targeted study recommendations (based on missed test items) is also provided. Detailed feedback includes (a) scores on each of the subscales within each module (three or four subscales are reported, depending on the content area), (b) a detailed listing and explanation of the knowledge, skills, and competencies demonstrated by the test-taker, and (c) a description of the knowledge, skills, and competencies that must be mastered to move to the next level of performance. This detailed information is provided both to test-takers who do not pass the module (to inform their study plans for preparation prior to retesting) as well as to test-takers who pass the module (as this information can be useful to them in better understanding the performance expectations that may underlie preparation for their educational or career goals).

Those persons who complete a GED Ready® exam are also provided detailed feedback. GED Ready® provides information on the types of items that were scored as incorrect. More specifically, the type of skill or knowledge necessary to answer an item correctly is provided, along with locations of cross-referenced publisher materials and resources that are specific to those skills and knowledge.

Additional information on score reporting is located in Chapter 6.

Chapter 2: GED® Test Validity Framework

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), define validity as “the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence for each interpretation is needed” (p. 225). Furthermore, the Standards define a validity argument as “an explicit justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores for their intended uses.” Explicit in these definitions is the need for a clearly defined test purpose statement as well as an explicit set of test score interpretations.

Kane (2013; 2006) provided a pragmatic framework for developing a validity argument. In this framework, the foundation of the validity argument is something called an *interpretation/use argument* (IUA), in which all the interpretations and uses of test results are clearly specified along with the logical flow of inferential statements and supporting assumptions.[5] The IUA involves identifying a series of if-then statements—each with its own set of assumptions—that build upon one another. Toulmin’s (1958) model is used to build each of these statements. The model involves a datum-warrant-claim progression in which some piece of information (datum) is used to make a claim. The warrant (and backing for the warrant) is used to support the claim. Each inference begins with a datum and ends with a claim. The claim becomes the datum in a subsequent if-then statement. The structure, then, becomes “if [datum], then [claim].”

Each inference is essentially a statement about the claims we make, given a piece of information. For example, we might make the following inference: if we can develop a set of items and tasks, then we can obtain an observed score. In this case, the datum is the set of items and tasks, and the claim is that we can obtain an observed score. The if-then statement is the framework for the inference, which could be stated as “the observed performances on items and tasks are evaluated to provide observed scores.” The observed score then becomes the datum in the next if-then inferential statement.

In practical terms, the IUA lists all the things that GED Testing Service has done, is doing, or will do in order to provide a test that serves a specific purpose statement. If the appropriate tasks have been conducted (with adequacy) and specified assumptions have been met, are defensible, or are otherwise reasonable, then an overall argument can be made regarding the validity of the test score interpretations and uses.

The validity argument itself is an evaluation of the accumulated evidence used to support the test score interpretations and uses. It is in this component that the IUA is evaluated. Kane (2006) states that “to claim that a proposed interpretation or use is valid is to claim that the IUA

[5] Kane (2013) recently renamed the interpretation/use argument (IUA) to emphasize both interpretations and uses. Formerly, the IUA was referred to as an interpretive argument.

is coherent, that its inferences are reasonable, and that its assumptions are plausible" (p. 23). The validity argument often involves the interpretation of empirical study results, but not every inference is necessarily supported by empirical evidence alone. Many assumptions can be supported by providing well-reasoned documentation. Regardless, each inference in the chain is evaluated as part of the validity argument.

GED® Test Interpretation/Use Argument

The following section provides a full description of an IUA for the GED® test. As noted in the test purpose statement, there are three purposes for the GED® test: (1) to provide a high school–level credential, (2) to provide information about a candidate's strengths and areas of developmental need, and (3) to provide evidence of readiness to enter workforce training programs or post-secondary education.[6]

Associated with each inference is a preliminary list of assumptions; these assumptions are subject to evaluation. For each assumption, we provide a brief description of the supportive evidence along with the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) addressed by each assumption, where applicable. These studies, when taken and interpreted together, weave the fabric of a strong, reliable, and evidence-based assessment system. These studies represent sound, constructive procedures that will be necessary for our assessment program.

It is extremely important to note that this set of projects is complementary and that evidence from any one project should not be interpreted independently. As suggested by the *Standards* (AERA, APA, NCME, 2014), the validity of test score interpretations is informed by a body of evidence collected via a continuous process. This set of studies is intended to provide that body of evidence.

Target Domain Inference

State CCR standards are extremely broad and defined across all years of K-12 education. The GED® test credential specifically emphasizes that test-takers demonstrate performance on high school–level standards. However, many of the high school standards rely on learning progressions and/or skills and knowledge acquired at earlier grades. Moreover, while we agree that all CCR standards are important and should receive emphasis within traditional high school settings, it is unrealistic to expect the GED® test to assess the full spectrum of CCR standards. Therefore, we must define a core set of academic knowledge and skills that an adult learner must demonstrate in order to be awarded a high school equivalency credential.

GED Testing Service began the test domain identification and development process in 2009. At that time, several states had already developed CCR curriculum standards, and the new state-driven Common Core State Standards (CCSS) were released for literacy and mathematics. All of these standards underwent a thorough review by GED Testing Service as well as myriad outside experts (see Chapter 1 of the Assessment Guide for Educators). Our staff worked

[6] Currently, the career- and college-readiness (CCR) indicator is planned as a post-operational launch scoring enhancement. Thus, the CCR indicator will not be part of the initial test scoring. Nevertheless, we have included the assumptions, inferences, and claims within the interpretation/use argument as a guideline for its development.

closely and extensively with internal and external content experts to identify those CCR instructional standards that were deemed “the core of the core” while simultaneously identifying the foundational standards from grades as early as grade 6.

The IUA for the target domain assumes that a clearly defined purpose exists for the test. The test purpose drives the test development process throughout. The purpose of the GED® test is documented here (see GED® Test Purpose section in Chapter 1) as well as the GED Testing Service website and throughout existing technical manuals and other documentation.

Another assumption associated with the target domain is that the instructional standards could be transformed into assessment targets, and that those assessment targets truly reflect the knowledge and skills associated with the instructional standards. Backing up even further, we assume that those who developed the CCR standards (both CCSS and other state standards) did so using relevant, appropriate, defensible, and accurate research.

The inference associated with the target domain is that representative, foundational CCR academic knowledge and skills can be identified from existing state-level K-12 instructional CCR standards. The standards that represent the core CCR knowledge and skills are adequate, given the purpose of the test.

Item and Task Development Inference

State CCR standards were written for instructional purposes and do not necessarily clarify how they are to be assessed. This fact means that the standards must be translated in such a way as to make them accessible for item development. To make item development possible, the instructional standards were translated into assessment targets, which in turn enabled GED Testing Service staff to develop fine-grained indicators for the assessment targets. Together, the assessment targets and their associated indicators were used to develop test and item specifications and blueprints.

At this stage, various decisions were made about the types of tasks and item formats that would elicit appropriate measurement of the assessment targets and indicators. These decisions were not minor ones, as item formats have serious implications on item/test development and scoring budgets.

Scoring Inference

The scoring inference is particularly complicated in that a broad range and number of steps are involved. To be clear, *scoring* refers to the steps ranging from assignment of numerical values to item/task performances to a final, reportable scale score. Specific steps may include assigning scores to items, summing over item scores to obtain unweighted and weighted raw scores, transforming raw scores to scale scores, and equating.

Items can be scored by the test delivery software or by humans, and sometimes by both. Each scoring method relies on different assumptions about the accuracy and appropriateness of the scoring method. Scoring details are too extensive to describe here but will be included in various technical manuals related to the GED® test.

GED Testing Service does not report raw scores. Instead, raw scores are transformed to a standard score scale for ease of interpretation. Standard scores are also associated with percentile ranks to aid in interpreting test performance relative to a nationally representative sample of graduating high school seniors. Given this transformation, we assume that the standard scores are assigned correctly, that rounding rules are appropriate and consistently applied, and that the standard scores have an appropriate range.

While the standard scores and percentile ranks are normative in reference, score reports also provide feedback regarding performance in content. Specifically, test-takers' scores fall into one of three performance zones (one zone that is below passing and two zones at passing and above) that span the reported standard score scale. Feedback is based on the typical performance of other test-takers whose scores also fell within that performance zone. We therefore make assumptions about the reference group from which "typical performance" is defined as well as the defensibility of the process itself (i.e., item-mapping process).

The scoring inference claims that we can obtain an observed score from the test items and tasks. The observed score serves as a basis for subsequent inferences and therefore deserves a good deal of attention with respect to issues such as reliability and accuracy.

Generalization Inference

Each test form is built to a set of test specifications in order to enable a standardized measurement. However, for obvious reasons (i.e., security) we have multiple forms for each test, and each form contains a differing set of items/tasks. The scoring inference claims that an observed score can summarize the test-taker's performance on a set of items/tasks. However, we are not interested so much about how well the test-taker performed on the specific set of items/tasks on a particular form. Instead, we want to make claims about their performance within the target domain.

To make this extrapolation to the target domain, we first have to infer that the observed score generalizes across various conditions of measurement, such as items/tasks, occasions, raters, and contexts. In theory, we would like to estimate the mean score for a test-taker across all possible conditions of measurement. Such a score is labeled a "universe score," and the observed score is our expected value of the universe score. Taking the mean over various conditions of measurement suggests that individual observations would vary across conditions. This variability is what we examine in reliability analyses.

The generalization inference infers that the items/tasks on any given form are a (random) sample of the universe of all possible items/tasks (also referred to as the universe of generalization in Generalizability Theory). We can be even more specific about the sample by looking at specific pieces of the test blueprint (e.g., content and/or cognitive specifications).

Extrapolation Inference

A test score must be representative of real-world applications. Thus, we extrapolate the universe score to the target domain. The *target domain* represents the extension of test performance to applications that may or may not have been included on the test itself. For example, the test may include reading comprehension applications within the context of a newspaper article. However, the test-taker may also be expected to apply reading comprehension skills to scientific journals or other texts, which may not be represented on the test. Furthermore, the application of the tested skills and knowledge occurs outside of standardized testing contexts, such as in a quiet room under timed conditions.

Decision Inference

Once we infer how a test-taker will perform on the target domain, we must subsequently make a decision about whether or not the person is qualified to earn a credential. From the target domain score, we infer that we can make a decision about the test-taker. The decision, in the case of the GED® test, is whether the test-taker has passed the content area test. A decision about passing the battery is made when all four content area tests have been passed. Additionally, there is a decision about whether the test-taker has met the criteria to be labeled as career- and college-ready.

Utilization and Consequential Inference

There are consequences associated with any decision made about the test-taker, based on the test performance. Regardless of whether or not the test-taker passes, there are ramifications. For example, if the test-taker does not pass, the consequence is that he or she may not be eligible for a job, promotion, post-secondary education, or job training program. The consequence for passing the test, however, is that the test-taker is likely eligible for many more opportunities. These consequences must be studied and assessed in order to make sure they are appropriate. For example, GED Testing Service should examine how employers, academic and vocational institutions, and the military utilize the GED® test and CCR credentials. Are they generally accepted as legitimate credentials, or do these entities require further verification or testing? Both positive and negative consequences must be studied in this context.

1. **Target Domain Identification and Development Inference.** Here we infer that a target domain representing foundational CCR knowledge and skills can be identified from existing, state-level, K-12 curriculum standards. The GED® test target domain is defined by the content standards and associated assessment targets. Each potential standard is a piece of information, or datum, and the claim is that the standards represent the foundational knowledge and skills that relate to the test purpose. The warrant for claiming that selected standards should be assessed is based on a decision about the necessity of the standards for demonstrating pre-specified levels of performance.

   The validity argument for this inference is based largely on logical arguments. Each of the assumptions can be supported by proper documentation of both process and decisions. For example, the standards that define the target domain are determined by groups of experts. Proper documentation would consider how the experts were selected by GED Testing Service, their qualifications, and their overall representativeness of the stakeholders (e.g., employers, college admissions officers).

Claim: A target domain representing foundational knowledge and skills associated with the test purpose can be identified.

Assumption	Supporting evidence	Type	Standard(s)
The test purpose statement is well articulated along with the intended score uses.	Documentation. Examinees and stakeholders understand the test purpose and intended score uses. Understanding the test purpose contributes to the identification of the target domain.	Logical	4.1; 8.1; 9.2; 9.7; 12.4; 11.13
Those who make recommendations regarding the target domain are representative of appropriate stakeholders (i.e., employers, college admissions officers, post-secondary instructors).	Provide evidence that experts are qualified and are representative of the field and stakeholders.	Logical	1.1; 1.9; 4.0; 4.6
Critical academic knowledge and skills can be identified using relevant research evidence.	Evidence is based on manner in which the Common Core State Standards and other non-CCSS states developed their curriculum standards. These standards are empirically linked to career- and college-readiness.	Logical	NA

Assumption	Supporting evidence	Type	Standard(s)
Standards can be ranked or ordered by importance such that core standards are identifiable. The selected standards represent the core academic skills and content knowledge.	Provide documentation of standards selection process. Provide documentation regarding justification for inclusion/exclusion of standards.	Logical	1.1; 4.0; 4.1; 4.2; 4.6; 4.12;
The target domain is compatible with the test purpose statement.	Provide documentation that the target domain adequately represents the skills and knowledge needed for career or college.	Logical	NA

2. **Item and Task Development Inference: From GED® target domain to observed performance on items and tasks**. If the target domain can be identified, then tasks and observable performances can be developed that measure (directly or indirectly) the target domain. Observed performances on the GED® test should reveal relevant academic knowledge and skills that represent the test purpose. The datum for this inference is the target domain. The claim is that the target domain can be translated into tasks that provide for measurable observations and performances. The warrant for the claim includes the test/item development specifications and test blueprint. Here again the majority of the evidence to support this inference is based on logical argument. Careful documentation of procedures, reasoning, and decisions should bolster the argument.

Claim: The target domain can be translated into tasks that provide for measurable observations and performances.

Assumption	Supporting evidence	Type	Standard(s)
Test/item development specifications and test blueprint are developed using relevant research evidence and with reason by experts who have adequate qualifications.	Test development documentation.	Logical	1.9; 4.2; 4.6
Assessment tasks that require important skills and content knowledge and are representative of the target domain can be identified.	Provide item development process documentation.	Logical	1.11; 4.0; 4.1; 4.7; 4.10; 4.12
Item types selected for use on the assessment support measurement of the specified target domain.	Provide item development process documentation.	Logical	1.11; 4.0; 4.1; 4.7; 4.10; 4.12
Items and tasks selected to stimulate responses can be mapped to different levels of the target domain/construct.	Provide item development process documentation.	Logical	1.11; 4.0; 4.1; 4.7; 4.10; 4.12
The test blueprint reflects the test purpose.	Test development documentation	Logical	4.1

3. **Scoring Inference: The observed performances on items/tasks are evaluated to provide observed scores.** If items and tasks can be developed, then we infer that an observed score can be obtained from the items and tasks. The datum for this claim is a candidate’s performance on a set of items or tasks, and the claim is that we can obtain a summative score. The warrant for the claim is the set of scoring rules. The evidence to support this inference will generally be established via typical operational activities. Field testing, for example, provides opportunities to determine whether scoring rules are accurate, appropriate, and consistent. There will be standard operating procedures regarding the selection, training, and calibration of raters that are intended to minimize rater effects.

Claim: An observed score can be obtained from the items and tasks.

Assumption	Supporting evidence	Type	Standard(s)
The scoring rules (scoring keys and rubrics) are appropriate.	Provide evidence to support scoring rules (e.g., why is one option correct while others are clearly incorrect?). Additional evidence may include the qualifications of those who develop the scoring rules. Provide evidence to justify and support the appropriateness of any rubric.	Logical	4.0; 4.7; 4.8; 4.18; 4.20; 4.21; 6.8; 6.9
Selection and training of scorers is adequate.	Provide evidence that essay readers meet or exceed documented qualifications. Provide evidence that readers have been and continue to undergo calibration checks.	Analytic/ Logical	1.9; 2.7; 4.0; 4.20; 4.21; 6.8
The scoring rule is applied accurately and consistently	Examine cognitive processes of human raters. Provide evidence that the rubric(s) result in a range of scores.	Analytic	2.7; 4.0; 4.13; 4.18; 4.19; 4.20; 4.21
The data fit the scaling model employed in scoring.	Examine dimensionality of test forms. Examine local item dependence. If LID is found, provide evidence that it is either non-systematic or that situation is rectified in an appropriate manner. Perform item fit analyses. Perform scale drift analyses.	Analytic	1.13; 4.8; 4.10; 4.13 4.20; 5.6

Assumption	Supporting evidence	Type	Standard(s)
All item types/formats can be scored accurately and consistently by automated systems.	Perform scoring key checks for machine-scored items. Perform scoring validation checks on regular basis.	Analytic	4.19; 6.8; 6.9
Automated scoring of short answer and extended response items adds value in terms of increased efficiency.	Explore and understand the uses and implications of automated scoring. Provide evidence that all tasks can be scored accurately, reliably, and efficiently.	Analytic/ Logical	4.0; 4.20; 4.21; 6.8
Items and tasks are developed without bias.	Facilitate Fairness Review Committees.	Analytic/ Logical	1.9; 4.0; 4.6
	Analyze item performance via field-testing.	Analytic	4.7; 4.10
	Conduct differential item functioning analyses, where possible.	Analytic	4.0; 4.1
Standard score scales are assigned correctly	Documentation of calculations, conversion tables, scoring verification procedures; rounding rules.	Analytic	Quality Control
Percentile ranks are assigned with accuracy	The sample from which the percentile ranks are derived is representative of the population of inference; sample size is adequate and timely.	Analytic	5.8; 5.9
Standard score scale has appropriate range	The standard score scale is carefully developed such that it adequately represents a range of observed scores.	Analytic	5.1; 5.2
Performance-level feedback is accurate	Item-mapping procedure; response probability thresholds Enough items to provide feedback	Analytic	1.14; 1.15; 12.19

4. **Generalization Inference: From observed score to universe score.** We infer that the items and tasks on the test are a random sample—or are at least a representative sample—of the test domain. The test domain represents the set of all possible observations across items, occasions, raters, etc. For example, we assume that all multiple-choice items on the Reasoning Through Language Arts test are interchangeable across forms, but each candidate is asked to respond to just one form (per test administration). The datum for this inference is an observed score and we claim that a candidate's observed score represents his or her expected score, or universe score, within the test domain. The warrant for the claim is that items or tasks are representative, or adequately sampled from the test domain (or universe of generalization). This inference essentially means that items should be interchangeable across forms, within the same content area. It should not matter which set of items or tasks each candidate receives. Support for this inference involves assessing the adequacy of any equating and the reliability and generalizability of test scores.

Claim: A candidate's observed score represents his or her expected score, or universe score, within the test domain.

Assumption	Supporting evidence	Type	Standard(s)
Test specifications and test blueprint provide for a test that clearly and accurately represents the test domain and allows for adequate forms selection.	Provide documentation of the test blueprint and specifications; ensure it is aligned with the test domain and avoids ambiguity during forms selection.	Logical	1.11; 4.1; 4.2; 4.12;
Test specifications are defined such that tasks and forms are parallel.	Provide evidence that equating procedures are appropriate. When standardization is not possible, such as when forms vary, statistical adjustments are used to equate scores. The justification of certain equating procedures requires support for the appropriateness of the equating model, the fit of the model to the data, and evidence indicating that the equating errors are not too large.	Analytic	5.6; 5.12-5.15
The observations made in testing are representative of the universe of observations.	Provide evidence that the test forms are aligned with the test blueprint.	Analytic/ Logical	2.9; 4.2; 4.12; 4.25
The sample of observations is large enough to control sampling/random error.	Provide evidence of reliability of test scores. Test scores should be generalizable over tasks. Evidence would include internal consistency, G-coefficient, or information functions.	Analytic	2.0-2.3; 2.5-2.7; 2.10; 2.13; 2.14; 2.16
Data collection occurred in a standardized manner, free of distractions and unusual events. Instructions to examinees are clear and comprehensive.	Provide evidence that standardization of data collection decreases variability. Standardization features, such as test format, instructions, and time limits, should decrease variability over data collections.	Logical	6.1; 6.3; 6.4

5. **Extrapolation Inference: From universe score to the level of content knowledge and skill (target score).** We infer that a universe score is representative of a person’s target score, or level of academic skill and content knowledge, as defined by the target domain. The target domain represents the range of all possible observations of a construct or trait (e.g., mathematics) and a person’s expected score is referred to as the *target score*. The target score is not observable. Increased levels of standardization in test administration result in a much narrower range of observations. For example, it would not be reasonable to expect our assessment to measure mathematical knowledge and skills in all possible contexts. We impose time limits on our tests, whereas in many real contexts this condition is not appropriate. This narrower range of observations is the test domain (see generalization inference above). The strength with which claims about competence in the target domain can be made will depend on overall confidence in the relationship between performance in the test domain and performance in the target domain. This will depend on how well the test tasks match the target domain activities and on any empirical or theoretical evidence for or against the linkage. The evidence to support this inference will come from logical argument, standard operating procedures, and additional empirical studies.

Claim: A candidate’s universe score represents his or her target score.

Assumption	Supporting evidence	Type	Standard(s)
The score earned on an assessment is indicative of an examinee’s level of performance in the target domain.	Conduct think-aloud and cognitive lab methodologies to determine whether candidates’ performances on tasks utilize the expected knowledge and skills necessary for job training programs or credit-bearing coursework in post-secondary education.	Analytic/ Logical	1.12
	Conduct content alignment studies to determine how, where, and to what extent our assessment aligns with other assessments that reportedly measure CCR. The study should evaluate the test purpose, content domain, and technical specifications.	Analytic/ Logical	1.16; 1.17; 1.18
The test tasks require the content knowledge and skills designated as necessary to support inferences regarding degree of performance on core foundational knowledge and skills.	Expert analysis; documentation.	Logical	1.11; 12.4

Assumption	Supporting evidence	Type	Standard(s)
There are no skills-irrelevant sources of variability that would seriously bias the interpretation of scores as measures of skill and content knowledge.	Provide evidence that accommodations result in comparable scores. Data from candidates with disabilities and others who receive accommodations or modifications should be analyzed to determine whether certain populations are disadvantaged in any form or manner.	Analytic	1.25; 3.0-3.6; 3.8-3.12
	Provide evidence that accommodations are administered correctly and fairly.	Logical	3.0-3.6; 3.8-3.12
	Provide evidence that extraneous factors did not interfere with examinee test performance. Studies should examine speededness, differential item functioning/bias, and reader scoring issues. Include testimony of participants, proctors, and administrators that nothing interfered with data collection.	Analytic/ Logical	1.25; 4.10; 4.13; 4.14; 9.13; 10.12
The items and tasks stimulate responses that indicate the full depth and breadth of the target domain to be measured.	Provide evidence that the items and tasks can be mapped to various assessment targets; assessment targets are fully represented among items and tasks.	Analytic/ Logical	1.11; 4.0; 4.2; 4.12

6. **Decision Inference: From conclusion about level of performance to scoring levels.** Using an estimated target score, we infer that a candidate's score can indicate demonstration of the foundational academic skills and knowledge necessary for high school exit as well as for career- and college-readiness. Specifically, we infer that candidates who do not meet the minimum benchmark for a high school credential have a need for additional development in foundational academic skills and content knowledge and are below the level of performance deemed necessary for high school exit. We infer that candidates who meet or exceed the high school exit benchmark (across all content areas) are at the level of proficiency in foundational academic skills and content knowledge consistent with that of traditional high school students upon graduation. Finally, we infer that candidates who obtain scores that meet or exceed the CCR benchmark (within a content area) demonstrate performance in the foundational skills and content knowledge at a level consistent with that needed to enter a workforce training program and/or a typical first-year, credit-bearing college course without remediation.

Claim: A candidate's score can indicate demonstration of the foundational academic skills and knowledge necessary for high school exit as well as for career- and college-readiness.

Assumption	Supporting evidence	Type	Standard(s)
Performance in the target domain depends on level of academic knowledge and skills measured by the test.	Evaluation of the alignment between skills assessed on the GED® test and the skills identified as necessary (or associated with) different types or categories of jobs. In order to provide potential employers with a complete profile of information regarding strengths and weakness, it is useful to relate performance on the GED® test to as many informative, external measures as possible.	Analytic/ Logical	1.5; 1.18
Decisions about candidates are made by assessing actual proficiency with the foundational academic knowledge and skills and not by candidate cheating or other test anomalies (e.g., nonstandard administration or using erroneous/incorrect instructions).	Test security is maintained; absence of evidence relating to test fraud. Regular scoring validation checks.	Analytic/ Logical	6.6; 7.9; 8.11; 13.8
The benchmark-setting process is clearly and explicitly defined before implementation.	Documentation	Logical	5.21; 2.22; 5.23

Assumption	Supporting evidence	Type	Standard(s)
The implementation of the procedures and data analysis is practical, as is the degree to which procedures are credible and interpretable to laypeople.	Documentation	Logical	6.10
Features of the Standardization and Norming study are reviewed and documented for evaluation purposes.	Documentation	Logical	5.2; 5.8; 5.9; 5.10; 5.21
The relationship between decisions made using the test and other criteria (e.g., grades, performance on a similar test, etc.) is known.	How much variation is explained by test scores, above and beyond what is explained by other measures?	Analytic	1.18
Decisions are reliable.	Reliability studies would include decision consistency and decision accuracy, conditional standard errors of measurement around the benchmark scores.	Analytic	14.15
Decisions have applicability.	Conduct a national survey of post-secondary institutions that are frequently targeted by GED® test credential holders. The purpose of the survey would be to collect information on course placement standards, such as minimum score criteria on assessments like ACCUPLACER. Subsequent to a linking study, this information would help GED Testing Service understand how GED® test scores relate to those standards.	Analytic	11.12
	Conduct a national survey of employers to determine GED® test graduates' readiness for employment. In addition, employers of GED® test credential holders could be asked to provide information on those employees regarding a range of characteristics and behaviors, such as job performance.	Analytic	11.12
The scores for the various content areas provide information to support decisions related to the purpose of the assessment.	Provide evidence that the tasks mirror the test domain. Judgment of experts stating that the tasks mirror the academic knowledge and skills required for readiness in job training programs or post-secondary general education courses and/or for high school exit.	Analytic/ Logical	1.11
Varying levels of success on the construct are specified and defined in a manner that supports the specified goals of the assessment.	Benchmarks fully describe the performance level of examinees. Reporting category scores also indicate areas of strength or developmental need.	Logical	5.21; 5.23

7. **Utilization/Consequential Inferences: From decision about performance classification to post-secondary admission and/or employability.** We infer that those who meet or surpass the GED® test credential have demonstrated a level of performance on academic knowledge and skills that makes them at least as academically qualified for careers and college as a typical high school graduate. We infer that those who surpass the career- and college-readiness benchmark will be ready for a job-training program or post-secondary education course.

Claim: GED® test credential recipients who score at or above the high school credential level are at least as academically qualified for careers and college as a typical high school graduate. Those who surpass the career- and college-readiness benchmark will be ready for a job, training program, or post-secondary education course.

Assumption	Supporting evidence	Type	Standard(s)
The GED® test scores are indicative of readiness for careers and college.	Acceptance of GED® test credential by employers, academic and vocational institutions, and military at a level consistent with that of high school graduates.	Analytic	1.5
The test results are used appropriately and are used to make defensible decisions about adults' readiness for jobs, job training programs, or entry-level, credit-bearing courses in post-secondary institutions.	Survey stakeholders regarding emerging, unintended, unanticipated consequences. It will be important to determine whether employers, educators, counselors, and so on are utilizing test scores in a manner consistent with the intended purpose of the assessment. For example, it would be inappropriate to associate GED® test scores with a grade point average. GED Testing Service has a responsibility to protect not only its brand but also its candidates, against inappropriate score usage.	Analytic	1.3; 1.15; 1.25; 5.3
Candidates receive usable information from the reporting system in order to improve learning and readiness.	Candidates who do not meet the high school credential benchmark benefit from additional instruction targeted to areas of developmental need.	Analytic	12.13

Assumption	Supporting evidence	Type	Standard(s)
Score reports provide adequate feedback, are interpretable, include valid information, and are timely.	Provide evidence that score reports are interpretable. Conduct usability studies with score reports to determine whether various stakeholders can interpret them correctly and efficiently. Users of score reports must be able to properly interpret test scores. Test scores should only be reported and interpreted within the context of the test purpose. Score reports must be carefully constructed so that only appropriate inferences are drawn from test scores.	Analytic/ Logical	5.0; 5.3; 6.10; 6.11; 7.10; 8.7; 9.16; 10.11; 10.12; 12.18
GED® candidate population does not change.	Evaluate changes in the GED® test-taking population. Given the overarching goal of the 21st Century Initiative ™ —to reach a larger population of GED test-takers—GED Testing Service should evaluate how much and in what manner the GED® test-taking population is changing from one year to the next, specifically with respect to size, demographics, age, and post-credential goals. In addition, surveys should be generated and distributed on an annual basis to support the interpretation of these results.	Analytic	9.7

Chapter 3: GED® Test Development, Specifications, and Forms Development

Assessment Development Process

The move toward career- and college-readiness (CCR) in both primary and secondary education puts those adults without a high school degree at further risk, not only in terms of lower post-secondary attendance and success but also in terms of their workforce readiness. As the knowledge and skills taught in the classroom evolve to meet CCR standards, adults without a high school diploma need a second opportunity to provide evidence that they are prepared to enter the workforce, post-secondary education, or both.

One of the major goals of the GED Testing Service in recent years has been the development of a new GED® test that indicates readiness for careers and college and continues to provide for issuance of a high-school equivalency credential. The philosophy underlying the new GED® test states that there is a foundational core, or domain, of academic skills and content knowledge that must be acquired for an adult to be prepared to enter a job, a training program, or an entry-level, credit-bearing post-secondary 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 for any post-secondary pursuit. An important goal of the movement towards career- and college-readiness content standards is that, as these new, more rigorous curricular standards gain ground in schools around the country, every high school graduate—and, by extension, every GED® credential holder–will possess the skills needed for success in the workforce or college. Today's reality, however, suggests that only a fraction of current high school graduates actually have such skills. Therefore, we continue to see great value in providing a second chance at attaining a high-school credential that is aligned with current expectations for high school graduates while still basing test content and test score reporting and feedback on a rigorous set of career- and college-ready expectations.

One important step toward achieving this goal was defining the targeted body of knowledge and skills within each content domain necessary to support these inferences. GED Testing Service consulted with a variety of advisory groups, expert panels, and stakeholder committees in order to identify and classify this core set of academic skills and abilities. The process resulted in the creation of Assessment Targets to support each of the four content area assessments of Reasoning Through Language Arts, Mathematical Reasoning, Science, and Social Studies.

The following discussion provides a description of the process implemented by the GED Testing Service to develop the Assessment Targets that define the focus of the new GED® assessments in Reasoning Through Language Arts (RLA), Mathematical Reasoning, Science, and Social Studies. The strategy that GED Testing Service and its collaboration partners adopted was to identify the core academic skills (across grades) considered to be most predictive of success in jobs, training programs, and credit-bearing post-secondary courses. Despite slight differences in the process used to develop Assessment Targets for Mathematics and RLA as compared to

Science and Social Studies, we accomplished this task using an iterative process that incorporated several stakeholder groups serving different roles throughout. The steps associated with this process are outlined in detail below.

Steps in the GED® Assessment Development Process:

1. **Definition of new GED® assessment goals**
   In order to appropriately define the range of knowledge and skills to be assessed with the new GED® test, the purpose of the assessment and the inferences it was intended to support needed to be clearly specified. To inform this task, we convened a panel of experts from a variety of different stakeholder groups to discuss the concept of "career- and college-readiness" and the manner in which it pertained to the GED® program's target test-taking population. This meeting was called the GED® Career- and College-Readiness Summit. Participants were provided with an introduction to the strategic direction of GED Testing Service, a brief overview of the preliminary assessment design (e.g., mode of administration or number of credential levels) and the overarching goals of the assessment (i.e., to provide for a high-school credential and support CCR inferences). They were then asked to provide comments on those characteristics they believed the new GED® test needed to possess in order to both achieve these goals and remain relevant and useful.

2. **Analysis of gap between existing GED® content specifications and the Career- and College-Readiness Standards**
   To gain a better understanding of the type and degree of change necessary to create an updated GED® test that would provide for inferences regarding CCR, we contracted with Achieve, Inc., to conduct a gap analysis. The gap analysis compared the content specifications that had been previously created for a (now discontinued) 5th Edition GED® test with the Common Core State Standards and other rigorous content standards (e.g., the National Assessment of Educational Progress standards and those used to define the domains assessed in particular Advanced Placement exams). This analysis helped identify areas in which we needed to increase rigor in order to create a test that would measure both readiness for high school exit and CCR.

3. **Establishment of the GED® Policy Board**
   A wide range of panelists representing organizations responsible for establishing educational policy was convened to provide comment on the specified goals of the GED® test, proposed test design, and the rationale behind the specifications described in an early draft of the Assessment Targets proposed by Student Achievement Partners (SAP), contributing authors of career- and college-readiness standards. This group helped us prioritize which skills and key concepts from the four content areas would be most versatile and useful for the unique GED® test-taking population.

4. **Drafting of Mathematical Reasoning and Reasoning Through Language Arts frameworks (RLA)**
   Once the high-level goals of the assessment were defined and the gap analysis complete, Achieve, Inc. convened panels of educational experts from high schools, 2-year and 4-year post-secondary institutions, and employers to identify the core competencies (across grade levels) that should be the focus of learning, instruction, and assessment in the content areas of Mathematical Reasoning and RLA.

5. **Definition of the assessment philosophy and Assessment Targets for Mathematical Reasoning and Reasoning Through Language Arts**
   Working with SAP, we used the Mathematical Reasoning and RLA content frameworks (developed in step 4) to identify a streamlined set of skills and body of content knowledge that became the draft GED® Mathematical Reasoning and RLA Assessment Targets. We employed data from research such as the ACT National Curriculum Survey to ensure that the skills selected were the most predictive of success in college and careers. This process also helped identify core literacy and quantitative reasoning skills appropriate for measurement on the Science and Social Studies tests.

6. **Development of indicators and refinement of draft Assessment Targets for Mathematical Reasoning and RLA**
   Because we needed the Assessment Targets not just to define the content domain of the new GED® test but also to provide more detailed guidance for test developers, we engaged content experts from Pearson School (PS) to help us define the individual skills to which items would be written. In a collaborative, iterative process, Pearson School, GED Testing Service, and SAP wrote, reviewed, and shaped several granular indicators that aligned to each overarching target in Mathematical Reasoning and RLA.

7. **Development of Science and Social Studies Practices**
   The literacy and quantitative reasoning skills defined by the career- and college-readiness Standards are intended to be versatile and useful across all content areas. Therefore, we were able to translate many of these key skills, along with some more content-area specific skills, into a set of practices that would be used to assess Science and Social Studies content. Like the Mathematical Reasoning and RLA targets and indicators, these practices were selected and revised collaboratively with SAP, PS, and GED Testing Service.

8. **Development of Content Topics for Science and Social Studies**
   In order to attain our goal of continuing to offer a high school–level credential, we recognized that GED® Science and Social Studies tests must define not only a set of skills relevant to these content areas but also a body of knowledge necessary for understanding basic concepts in the sciences and the social sciences. For this purpose, GED Testing Service and Pearson School selected a team of content experts to develop a set of content topics that would provide context for each item on the Science and Social Studies tests. These content topics were then reviewed, revised, and categorized in a collaborative process with SAP, GED Testing Service, and PS.

9. **Review of assessment targets by adult educators**
   GED Testing Service convened panels of adult educators and state adult education program directors to focus on particular content areas. These panels reviewed the targets, indicators, and content topics for clarity and completeness. GED Testing Service and Pearson School collaboratively made revisions based on the panels' feedback to ensure that items written to the Assessment Targets would be relevant to the needs of adult learners.

10. **Generation of test blueprint documents**
    Once Assessment Targets for all four content areas were finalized, content experts from PS and GED Testing Service grouped indicators in a variety of different ways in order to ensure appropriate distribution of content across test forms. Blueprint specifications were determined by a number of content parameters that appear at the beginning of each Assessment Target document. These parameters constitute guidelines for item distribution across Depth of Knowledge[7] (cognitive complexity) levels, item distribution across the indicators, and several other considerations.

Many competing desires and constraints had impacts upon both the Assessment Targets and the test design itself. One such constraint was the time given to test-takers to complete the entire test battery. The remainder of this section takes a top-down approach to describing the processes used to develop both an overarching philosophy behind the new GED® test and a detailed set of content specifications that were used to guide initial and current test development. It also discusses how the many competing requirements were addressed to meet the purpose of the assessment.

The following four parts will discuss the Assessment Target development process in greater detail. Because many of the steps delineated above occurred in an overlapping and recursive manner, we begin with the highest-level considerations and then drill down through successively smaller issues.

Part 1: Establishing Goals and Purposes for the New GED® Test

In the early stages of identifying the domain of academic skills and knowledge to be assessed on the new GED® test, GED Testing Service convened a series of meetings with thought leaders from various education, employment, and assessment communities. The first of these panels, the Advisory Board, met at Career-and College-Readiness Summits in January 2011 and again in March 2011 to determine (a) what the term "career- and college-readiness" (CCR) means to the GED® test-taking population and b) what information college admissions, employers, instructors, and test-takers themselves most needed from a high-school credential and a CCR credential. The second panel, the GED® Policy Board, met in May 2011 to discuss a variety of issues central to the GED® program, including the breadth of content to be assessed,

[7] The Depth of Knowledge (DOK) model is a framework for analyzing a wide range of educational materials on the basis of the cognitive demands they require in order for a learner to produce a response to those materials. In 1997, Dr. Norman L. Webb developed this model for analyzing the varying levels of cognitive complexity in academic standards and curricula. For more information, see http://www.aps.edu/rda/documents/resources/Webbs_DOK_Guide.pdf

how to position the new test in the market, and how to build content validity around a dual-credentialing system.

Participants on both the Advisory Board and the Policy Board were strategically selected in order to ensure that we received input from as many key stakeholder groups as possible. Both of these panels were designed to be broadly representative of potential end users of the new GED® test. The Advisory Board included representatives from the adult basic education community and other key educational organizations, such as the Council of Chief State School Officers. We also invited leaders from workplace organizations, such as The Manufacturing Institute and the Department of Defense, who could potentially offer viable pathways to GED® credential holders. In addition, experts in large-scale standardized assessment and educational policy played key roles in the discussion. The Policy Board, however, required a slightly different group of experts. For this convening, we invited a number of state- and district-level education policy-makers from key GED® jurisdictions, such as Florida, New York, and Michigan. We also sought input from other large-scale assessment programs, such as the ACT and Smarter Balanced Assessment Consortium, as well as from the secondary and post-secondary education communities. This wide array of leaders from all levels of education and the workplace contributed to shaping the new GED® test design from a very high level.

Career- and College-Readiness Summit

In collaboration with Achieve, Inc., GED Testing Service hosted the January 2011 Summit in order to receive input from the GED CCR Advisory Board on defining CCR for the GED® population. The input provided helped GED Testing Service think about the types of credentials to be awarded to GED® test-takers and how they should be operationalized. The panel considered that the movement toward more rigorous national standards for high school education, as well as the changing national conversation about what is needed to be ready for college and careers, had a direct impact on the GED® testing program. The panel agreed that the GED® testing program must be re-envisioned in order to align itself with this new direction. Although panel members understood that the Career- and College-Readiness Standards would serve as the benchmarks for the GED® assessment frameworks, one of their tasks was to consider which CCR academic skills they would most value in potential GED® credential holders, given the unique characteristics of the wide range of adult learners who comprise the GED® test-taking population.

The first CCR Summit primarily focused on issues of test design and credentialing, and the second CCR Summit allowed GED Testing Service to update the Advisory Board on its progress. For the purposes of elucidating this aspect of the Assessment Target development process, we focus here on the first CCR Summit.

In order to help focus the discussion at the January 2011 Summit, we posed the following questions on issues pertaining to test design and development to the panel:

1. Since 1944, the GED® credential has been the only credential to confer basic high school equivalency. However, with the current national emphasis on college- and career-readiness, the current GED® credential has the potential to lose relevance and credibility. Other credentials that provide specific information about career readiness, such as WorkKeys, have emerged to address users' needs in recent years. How is the current GED® test falling short of serving admissions counselors, training programs, and employers?
2. Users of GED® test results include post-secondary admissions counselors, training programs, and employers. What do these users need to know to make decisions about a GED candidate's preparedness for post-secondary admissions, career training programs, and/or readiness for a specific job? How could a new GED® credential add value in decision-making?
3. How should the GED® credential be structured in order to preserve its value as the sole source for American high-school equivalency while serving other critical purposes?

Several key points emerged from this discussion related to the intended use of the results.

1. Measuring a well-defined set of core competencies in the content domains of mathematics and literacy is of primary importance to end users, including admissions counselors, employers, and test-takers. All three groups of stakeholders tend to value a similar set of core skills.
2. There is support for a dual-level credentialing system that would provide more specific information about what GED® credential holders know and can do.
3. It is recommended that the new test provide an opportunity for test-takers to demonstrate basic competence with technology and computer navigation skills.

The Policy Board Meeting

The Policy Board, which convened in May 2011, served a slightly different function in defining the new GED® test system from a high level. As we had already begun development of the content frameworks (see Part 2 of this section), we were able to solicit feedback from policymakers and education leaders about both the overall design of the test and the priorities for test content that were established in early drafts of the Assessment Targets for Mathematics and RLA. The Policy Board also helped determine some priorities for what content would be included in the Science and Social Studies Assessment Targets.

The following key questions helped guide the Policy Board's discussion:

1. Our goal is to develop assessment targets that focus students and teachers on a powerful skill set that they can deploy across the subject areas and predict readiness in career, college, and life. Do the assessment target drafts [that panel members were provided] for RLA and Mathematics reflect this goal?
    a. What critical knowledge and skills are missing?

   b. Should any assessment targets be taken off?
2. What focus areas for Science and Social Studies should be added or deleted to make those tests reflective of the knowledge and skills that adults should have in order to predict readiness in career, college, and life?

The Policy Board's key findings and recommendations included the following:

1. The focus in the RLA targets on reading complex texts drawn from social studies and science subjects as well as workplace documents is appropriate and necessary for the GED® population. Similarly, the focus on core computational competency and depth rather than breadth in the Mathematics targets is applicable and essential for GED® test-takers entering into a wide range of post-testing pathways.
2. The panel recommended that the Science and Social Studies tests require test-takers to demonstrate not just the ability to answer text-based questions about the content but also apply their knowledge of important key concepts in the sciences and social sciences.
3. The panel also acknowledged that because approximately only 40percent of high school graduates are currently prepared to enter a credit-bearing post-secondary course or job training program, GED Testing Service's approach of continuing to offer a high school–level credential in addition to a CCR endorsement is reasonable and appropriate.

Through this succession of meetings with diverse groups of experts and stakeholders in the GED® program, a new vision of both the role of the GED® test as part of the adult educational system in the United States began to solidify. Additionally, the content domains in each content area began to achieve better specificity and focus.

Part 2: Drafting New GED® Content Frameworks and Aligning to Career- and College-Readiness Standards

The Assessment Targets for the new GED® test need to serve multiple purposes and multiple audiences. They must draw a clear connection between the skills measured in the assessment and national career- and college-readiness standards, such as the CCSS. They must provide detailed guidance to test developers about the depth and breadth of content covered. They must clearly define the domain of the GED® test so that instructors and curriculum designers can develop appropriate materials for adult learners with a wide range of skill levels. Additionally, they must contribute clear, consistent evidence that supports the validity of the inferences to be made with assessment results.

In order to begin this work, we enlisted the help of Achieve, Inc. to conduct a gap analysis between CCR Standards and content specifications that had been created to guide the development of a "5th Edition" GED® test. The 5th Edition test, originally slated to launch in 2012, was to be a moderate revision and update to the current test series, which was launched in 2002. However, Achieve's gap analysis pointed to a number of significant differences between the proposed 5th Edition test and the expectations associated with students finishing high school in a CCR-based curriculum. Therefore, this analysis led to a recognition that what was required for the next-generation GED® test was not simply an update but rather a thorough and complete re-envisioning of the test itself and the various adult education programs that support it.

Key findings from the gap analysis included the following:

1. Content specifications for all four content areas should provide greater clarity for both test developers and instructors. They should present a conceptual framework that offers a pathway to measuring those skills considered most important for test-takers seeking a high school–level credential to possess.
2. Though the GED® content specifications themselves demonstrated some overlap with nationally recognized content standards, such as the National Assessment of Educational Progress (NAEP) objectives and various CCR standards, sample items from the GED® 2002 Series (the test in circulation at the time of this publication) provided to Achieve did not measure the skills with a level of rigor appropriate for a high school level.
3. Although the specifications for the GED® Science and Social Studies tests required test-takers to demonstrate proficiency with basic content-area literacy skills, they did not make it necessary for test-takers to understand important concepts in the various disciplines within the sciences and social sciences.

In response to these findings, we expanded our collaboration with Achieve to include drafting content frameworks for the new test that would be derived from CCR standards. We knew at the onset that this drafting process would be iterative and would require input from a number of content experts. We also knew that the primary goal for the early drafts was simply to identify a streamlined yet elegant and powerful set of skills from within the CCR Standards. These skills would need to be (a) relevant, useful, and versatile for our especially diverse testing population and (b) able to be reliably measured in a proctored and timed large-scale assessment format.

In January and February 2011, Achieve convened two panels of content experts, one for Mathematical Reasoning and one for Reasoning Through Language Arts. The participants in these panels represented high school educators, 2- and 4-year post-secondary institutions, and workplace education experts. Additionally, some panel members had served as officers in national teachers' associations such as the National Council of Teachers of Mathematics (NCTM) and the National Council of Teachers of English (NCTE) as well as educators who had participated in reviews of the CCR standards. Each of these panel members was selected to broadly and generally represent the various pathways available to GED® credential holders and to provide content expertise.

The process for identifying key skills for inclusion on the new GED® test involved a careful analysis of each CCR standard. For the English language arts standards, each of the learning progressions were examined to determine which individual skills were essential for instruction and which skills were essential for both instruction and assessment. Similarly, the mathematics standards were evaluated on the criteria of their utility and versatility. Some of the mathematics skills were understood to be foundations necessary for understanding higher-order quantitative reasoning. Many of these skills were identified as having great importance to the GED® test-taking population and are therefore essential for both instruction and assessment.

Achieve compiled the data from the subject matter expert analysis of the CCR standards into a framework draft document that served as the basis of the Assessment Targets. This document contained a summary of mathematics and literacy skills and categorized them each as “instructional” or “essential for assessment.” Test developers at GED Testing Service then further analyzed the skills identified as “essential for assessment” to verify that these skills could indeed be measured reliably in a standardized environment. For example, the English language arts panel recommended that we create a performance task that allowed test-takers to demonstrate their speaking and listening skills. However, because the technological and cost constraints about recording and scoring these types of responses are significant, we made strategic decisions to reserve speaking and listening tasks for the long-term goals for the new GED® test.

The framework draft that emerged from the collaboration with Achieve and the subject matter experts contained a comprehensive summary of the CCR standards and therefore provided for a larger universe of content than could be reliably measured given the time constraints of the new GED® test. Therefore, we began our collaboration with SAP. Through this collaboration, we were able to begin to streamline the frameworks in order to identify specific content targets that could be assessed.

Through an iterative review process, a streamlined set of skill categories began to emerge in both Mathematics and RLA content areas. In RLA, we were able to use the CCR Reading, Writing and Language Anchor Standards in order to distill the learning progressions that span multiple grades into a clearly articulated set of skills considered most relevant to GED® test-takers. During these discussions, we also determined that the cognitive complexity and difficulty of most items on the RLA test would be driven by the textual complexity of their associated passages. Therefore, the passages would be selected to provide a range of text complexity levels. Items written to this range of passages would then provide an opportunity for us to gather information about test-takers who had a wide range of abilities.

For Mathematics, in which the learning progressions are not as clearly delineated, we identified a number of core concepts necessary for quantitative and algebraic reasoning. Many of these core skills are introduced in the CCR standards in the early grades, but they are successively reinforced as students progress through their coursework to ensure deep, high-school–level mastery. The collaborative drafting process with Student Achievement Partners allowed us to consider how best to target these skills in a way that would support test-takers in developing a strong foundation on which they could build higher-order Mathematical skills. We also considered how to assess test-takers' ability to employ these foundational skills at a range of ability levels. On the basis of these considerations, we developed the Mathematics Content Matrix in order to clearly and graphically represent a proportional breakdown of the mathematics content skills per test form.

Part 3: Creating Item-Level Skill Descriptors

As part of the Assessment Target drafting process, we recognized that we would need to provide specific guidelines to test developers and item writers regarding the content that individual items would measure. Therefore, concurrently with the development of the targets, GED Testing Service began to work with PS to draft the more granular skill descriptors that we call “indicators.”

The process used to compose the indicators included distilling the essential skills from progressions of standards from across grade levels in the CCR Standards. We took the basis of each indicator from the grade level in which the foundational skill was most clearly stated and supported the skill description with important aspects captured in other grade levels.

Once we had drafted these indicators and ensured that they were each aligned with a specific target, various consultants, including SAP, reviewed and edited the targets and indicators into a single document per content area. Once again, this review process was iterative and resulted in a set of clearly articulated and distinct skills that would inform both instruction and item development.

Toward the end of the iterative reviewing process of the indicators, GED Testing Service convened a panel of adult educators and adult education program directors from around the country. In these meetings, representative members of the adult education community provided feedback regarding the relevance of the Assessment Targets for Mathematics and RLA to their students. The input from this group was crucial for finalizing which skills were ultimately to be assessed on the new GED® test. With this final step completed, we were ready to publicly release the Assessment Targets.

Part 4: Science and Social Studies Practices and Content Topics

Although the development processes of the Assessment Targets for Science and Social Studies were certainly similar in some ways, the documents for these two content areas called for a somewhat different set of requirements. Because we are committed to assessing not just the skills required for career- and college-readiness but also those broadly and generally representative of what is taught in high schools across the United States, we understood that we needed to define the domain of science and social studies content knowledge that GED® test-takers must know.

First, we defined the structure of the Science and Social Studies Assessment Targets slightly differently from the Mathematics and RLA Assessment Targets. Instead of listing a set of essential skills, we created a two-part document that listed both skills and content topics. The skills section became the Science or Social Studies Practices. To define the practices, we consulted several existing sets of nationally recognized content standards. From various CCR standards, we identified a selection of literacy and mathematical skills that had particular application in the sciences and the social sciences. For example, we focused on data representation and descriptive statistics skills that would require test-takers to glean information from tables, graphs, or maps in scientific or historical contexts. Similarly, we created practices around the skills of being able to read complex scientific texts as well as primary and secondary source documents.

We also identified a number of content-area-specific skills that are key to successfully understanding scientific and historical content. These skills include concepts such as designing experiments and testing hypotheses in science and distinguishing between supported and unsupported claims in social studies. For the basis of the science practices, we consulted the National Research Council's "A Framework for K-12 Science Education: Practices, Crosscutting Concepts and Core Ideas,"(2011) and for social studies, we consulted the NCSS National Curriculum Framework for Social Studies (2010) and the National Standards for History, Revised Edition (1996).

Following a similar process as that established during the development of the Mathematics and RLA Assessment Targets, SAP, GED Testing Service, and PS reviewed and revised the practices though a series of iterations. Simultaneously, GED Testing Service contracted with teams of content experts in science and social studies to compile a list of the topics and key concepts that define the domain of these content areas. In social studies, we worked with experts in civics and government, United States history, economics and geography, and in science, we worked with experts in biology, chemistry, physics, and earth sciences. These experts, many of whom had also contributed to assessment frameworks such as those developed for the NAEP and other tests, proposed an extensive list of topics that are typically covered in a four-year high school course of study.

This list was lengthy and significantly detailed. Because we have limited testing time and a relatively small number of items on each of the Science and Social Studies tests, we consolidated many of the topics listed so that we could (a) create "subtopics" of consistent grain size and b) eliminate some topics that fell beyond the bounds of the new GED® content domains. Therefore, in collaboration with Student Achievement Partners and Pearson School, we began paring down the list of content topics in two ways. First, we grouped related ideas together and collapsed many of the details into single subtopics. Then we identified "focusing themes" that would further limit the content that would be assessed. These themes were selected to ensure the tests would cover a wide range of important concepts and ideas in science and social studies but also function like a lens in that they draw focus to a distinct subset of ideas within each content topic. That is, each content topic contains a broad group of ideas; however, each topic is aligned with a particular theme, and content that falls outside the parameters of the theme will not be covered on the tests.

The two focusing themes for Science are as follows:

> **Human Health and Living Systems.** This theme pertains to material that is vital for health and safety of all living things in the modern world. Topics that may be explored from this area of focus include the physical body and characteristics of humans and other living things, their systems as well as topics like diseases, evolution, and heredity. This crosscutting concept may also delve into the mechanisms for how the human body works on a chemical and physical level within the domain of physical science. Within

Earth and space science, topics come from how the environment affects living things and human society as well as how humans and other organisms affect the environment.

Energy and Related Systems. This theme deals with a fundamental part of the universe. Topics cover everything from sources of energy, transformations of energy, and uses of energy. Within the domain of life science, this theme is reflected in content exploring how energy flows through organisms and ecosystems. Similarly, the Earth's geochemical systems are touched upon in Earth and space science. Topics related to how humans gain energy in their bodies and the results of the use of that energy are also relevant to this theme.

The two focusing themes for Social Studies are as follows:

Development of Modern Liberties and Democracy. This theme explores the development, from ancient civilizations to the present, of current ideas about democracy and human and civil rights. It examines contemporary thinking, policies and structures, major events that have shaped our democratic values, and major thinkers who contributed to American ideas of democratic government.

Dynamic Responses in Societal Systems. This theme explores how the systems, structures, and policies that people have created respond to each other, conditions, and events. For example, societies and civilizations have developed and changed in response to particular geographic features and natural events. National economies respond to both governmental policies and natural laws of economics, such as supply and demand, around which policies are built. Similarly, countries have responded to changes and challenges—both internal and external—in ways that were beyond the ability of any one person to control.

Once the focusing themes were established, we were able to fine-tune the content topics to ensure that they each corresponded with a theme and could be interpreted narrowly within the confines of the theme. In this way, the themes have allowed us to limit the universe of content that is assessed on the new GED® test to that which is most relevant, versatile, and useful for our test-takers.

With the two major components of the practices and the content topics complete, we were able to bring the Science and Social Studies Assessment Targets to the panel of adult education reviewers for review and feedback. As with Mathematics and RLA, this panel evaluated the targets to ensure that described content would be relevant to their students and provided editorial suggestions that allowed us to finalize the documents.

Part 5: Creating Test Blueprints from the Assessment Targets

The new GED® Assessment Targets for the four content areas of Mathematical Reasoning, Reasoning Through Language Arts, Science, and Social Studies serve several audiences and purposes. Though they undoubtedly have an impact on adult learners and adult education instructors, they are primarily intended to provide guidelines to test developers who build the new GED® test based on the parameters laid out therein. As such, the Assessment Targets also include several parameters that have helped shape test blueprints.

These parameters each provide the basis for one “layer” or dimension of the test blueprint. For instance, the following is the list of content parameters for the Social Studies test (parameters for each content area vary slightly but essentially follow this model):

1. Roughly 50 percent of the test focuses on civics and government, roughly 20 percent focuses on United States history, roughly 15 percent focuses on economics, and roughly 15 percent focuses on geography and the world.
2. The test includes items that assess textual analysis and understanding, data representation and inference skills, and problem solving with social studies content.
3. Each item on the Social Studies Test is aligned to both one Social Studies Practice and one Content Topic.
4. Each item is also aligned to one Depth of Knowledge level of cognitive complexity, based on the appropriate alignment to a Social Studies Practice.
5. Roughly 80 percent of the items are written to Depth of Knowledge level 2 or higher.
6. The contexts within which problem-solving skills are measured are taken from both academic and workforce contexts.
7. Roughly 50 percent of the items are presented in item scenarios, in which a single stimulus (which may be textual, graphic, or a combination of both) serves to inform two to three items. The rest of the items are discrete.

In following these parameters, test developers can ensure appropriate coverage of important aspects of the content on each individual test form. As per the above, each Social Studies form includes a spread of items showcasing an array of content, Depth of Knowledge levels, and item types that is appropriate for measuring the targeted domain.

GED® Test Specifications for all Content Areas

Item Types

The variety of item types available for use on the GED® test is larger now, thanks to computer-based testing. The computer-based testing platform gives the opportunity to use interactive item types that are not possible on a pencil-and-paper test. Each content area test features an assortment of item types listed below, some that already appear on the 2002 Series GED® Test, and others that are new.

Item Types in Reasoning Through Language Arts

The GED® RLA Test on the new assessment is composed of several passage sets. Each passage set includes text ranging from 400 to 900 words and six to eight items. The RLA Test features the following:

- Multiple choice items
- Several different types of technology-enhanced items
- Drop-down items embedded in passages
- One 45-minute extended response item

These items assess the full depth and breadth of skills outlined in the GED® RLA Assessment Targets. Test-takers can apply different cognitive strategies with the wide variety of item types, demonstrating proficiency with the RLA content. This allows GED Testing Service to assess the targeted content at a number of Depth of Knowledge (DOK) levels.

Multiple choice (MC) items are used to assess aspects of virtually every indicator listed in the GED® RLA Assessment Targets. This item type continues to be a reliable method for measuring skills and knowledge at a range of cognitive levels in a standardized manner. Unlike the MC items on the 2002 Series GED® Test, the MC items on this new assessment only have four answer options, rather than five. This is the only content-area test in which each MC item refers to a passage.

Fill-in-the-blank (FIB) items can also be used to measure a wide range of skills identified in the GED® RLA Assessment Targets. In particular, they may provide the unique opportunity to assess vocabulary skills at a higher cognitive level than MC items might by requiring test-takers to supply their own synonyms rather than choosing from four options. FIB items require test-takers to create a short phrase or complete a sentence in order to analyze a text feature within a passage.

Drag-and-drop items are interactive tasks that require test-takers to move small images, words, or short phrases to designated drop targets on a computer screen. They are often used to assess a test-taker’s ability to classify and appropriately sequence information. For example, a drag-and-drop task might require test-takers to order events in a passage on the basis of chronology or cause and effect. They may also provide opportunities for test-takers to analyze an author’s arguments by classifying the arguments as true or false. These items may employ a variety of different graphic representations, including Venn diagrams, timelines, and many others. Another way the new GED® assessment may employ the drag-and-drop technology is in editing tasks that require test-takers to reorder paragraphs within a passage or sentences within a paragraph.

Drop-down items are items with multiple response options embedded directly within a text. On the RLA Test, this item type is used primarily to assess the language skills, such as conventions of Edited American English, standard usage, and punctuation, outlined in the GED® RLA Assessment Targets. These items are designed to mimic the editing process as authentically as possible; therefore, variations of a phrase appear as options in drop-down menus within the text. Once the test-taker selects an option, the answer shows on the screen as part of the text.

The Extended response (ER) item on the RLA Test is a 45-minute task that requires test-takers to analyze one or more source texts in order to produce a writing sample. The source texts do not exceed 650 words. These ERs are scored on three dimensions, as outlined in the Extended Response Multi-trait Scoring Rubric. The first trait on the rubric pertains to how well test-takers analyze arguments and gather evidence found in source texts in support of the positions that they take in their writing samples. The second trait scores the writing samples on the basis of how well the writing is developed and organized.

The writing samples are also scored for how well test-takers demonstrate fluency with conventions of Edited American English, per the third trait on the rubric. Each of these three traits will be scored on a four-point scale. The prompts for the ERs are developed to elicit analytic writing that effectively uses evidence from the source text(s).

Item Types in Mathematical Reasoning

The new GED® Mathematical Reasoning Test features the following:

- Multiple choice items
- A variety of technology-enhanced item types
- Drop-down items

These items assess the full depth and breadth of skills outlined in the GED® Mathematics Assessment Targets. Employing a wide variety of item types also allows us to assess the targeted content at a number of Depth of Knowledge (DOK) levels, as they each provide opportunities for test-takers to apply different cognitive strategies to demonstrate proficiency with Mathematics Test content. Each item type on the Mathematics test may be presented either as a discrete item or as part of an item scenario in which two or three items pertain to a single stimulus. Stimulus materials may include brief text, graphs, tables, or other graphic representations of numeric, geometric, statistical, or algebraic concepts.

Multiple choice (MC) items are used to assess aspects of virtually every indicator listed in the GED® Mathematics Assessment Targets. This item type continues to be a reliable, standardized method for measuring skills and knowledge at a range of cognitive levels. Unlike the multiple choice items on the 2002 Series GED® Test, the MC items on the new assessment have four answer options rather than five.

Fill-in-the-blank (FIB) functionality on the Mathematics Test gives the test-taker the opportunity to type in the numerical answer to a problem or to enter an equation using keyboard symbols or the character selector.

Drop-down items with drop-down menu functionality are used to give test-takers opportunities to choose the correct math vocabulary or numerical value to complete statements. As with editing tasks in the RLA Test, the test-taker is given the advantage of seeing the complete statements on screen in an authentic way. Drop-down items are frequently also used to make comparisons between two quantities.

Hot spot items consist of a graphic image with virtual "sensors" placed strategically within the image. This item type can be used to measure skills with regard to plotting points on coordinate grids, on number lines, or on dot plots. Test-takers can also select numbers or figures that have a particular characteristic or create models that match given criteria (e.g. given a three-dimensional figure, the test-taker could select its edge or create a model of two-thirds of a rectangle divided into 15 sections). Hot-spot items create a much more authentic experience for test-takers because they provide opportunities for test-takers to navigate within a two-dimensional field to demonstrate their proficiency with a variety of quantitative, algebraic, and geometric skills.

Drag-and-drop items are interactive tasks that require test-takers to move small images, words, or numerical expressions to designated drop targets on a computer screen. They can be used to create expressions, equations, and inequalities by dragging numbers, operators, and variables into boxes that form an equation. Drag-and-drop items can also be employed in the service of demonstrating classification and sorting skills, as they provide an opportunity for test-takers to organize data based on a set of characteristics. The test-taker can also order steps in a process or solution or match items from two sets.

Item Types in Science

The new GED® Science Test features the following:

- Multiple choice items
- Short answer items
- A variety of technology-enhanced items
- Drop-down items

These items assess the full depth and breadth of skills outlined in the GED® Science Assessment Targets. Employing this variety of item types also allows us to assess the targeted content at a number of Depth of Knowledge (DOK) levels. Each item type provides opportunities for test- takers to apply different cognitive strategies to demonstrate proficiency with Science practices and content knowledge. Each item type on the Science Test may be presented either as a discrete item or as part of an item scenario in which two or three items pertain to a single stimulus. Stimulus materials may include brief text, graphs, tables, or other graphic representations of data or scientific concepts. Many of the Science Test stimuli pertain to the focusing themes of “Human Health and Living Systems” and “Energy and Related Systems,” as identified in the GED® Science Assessment Targets.

Multiple choice (MC) items are used to assess aspects of virtually every Science Practice and Content Topic listed in the GED® Science Assessment Targets. This item type continues to be a reliable, standardized method for measuring skills and knowledge at a range of cognitive levels. Unlike the multiple choice items on the 2002 Series GED® Test, the MC items on the new assessment only have four answer options rather than five.

Fill-in-the-blank (FIB) functionality on the Science Test gives a test-taker the opportunity to type in the correct response when potential answers have little variability. For example, this item type can be used when an item calls for a response to a specific calculation or when the test-taker is required to excerpt a word or phrase from a text to demonstrate understanding of an idea or vocabulary term. More specifically, a particular item measuring data interpretation skills in a science context could call for a single word or short phrase to describe a trend on a graph.

Drop-down items with drop-down menu functionality embedded within a brief text are used to give test-takers opportunities to choose the correct response to complete statements. As with editing tasks in the RLA Test, test-takers are given the advantage of seeing the complete statements they create in an interactive manner onscreen. These items can measure many of the same skills that fill-in-the-blank items can, though they provide a selection of possible responses from which test-takers can choose.

Drag-and-drop items are another type of interactive task that require test-takers to move small images, words, numerical expressions to designated drop targets on a computer screen. On the Science Test, this item type can be used to measure a test-taker’s skills with regard to assembling data or comparing and classifying information. For instance, an item could ask test-takers to place organisms in specific locations on a food web. Other examples of tasks well suited to drag-and-drop items might be ones in which test-takers place labels on a graph or chart, fill in a Venn diagram with data from a brief textual stimulus, order steps in a scientific experiment, or place data points from a given context into a chart, table, or graphical model.

Hot spot items consist of a graphic image with virtual “sensors” placed strategically within the image. They can be used to measure a test-taker’s understanding of relationships between data points cited from a textual or graphic stimulus. For example, a hot spot item could contain a pedigree chart requiring test-takers to select offspring with a particular trait in order to demonstrate their understanding of heredity. Other items might ask test-takers to select data or points in a graph, chart, or table that support or refute a given conclusion or to select parts of a specific model given some selection criteria (e.g., a model of the human body, a cladogram, or a matter-cycle diagram).

Short answer (SA) items provide opportunities for test-takers to demonstrate a wide range of cognitive strategies as they compose their own brief responses to the wide range of content outlined in the GED® Science Assessment Targets. This item type could be employed to determine whether a test-taker can provide a valid summary of a passage or model, create and successfully communicate a valid conclusion or hypothesis, or derive evidence from a textual or graphic stimulus that specifically and accurately supports a particular conclusion.

Item Types in Social Studies

The new GED® Social Studies Test features the following:

- Multiple choice items
- A variety of technology-enhanced items
- Drop-down items
- One 25-minute extended response item

These items assess the full depth and breadth of skills outlined in the GED® Social Studies Assessment Targets. Employing this variety of item types should also allow us to assess the targeted content at a number of Depth of Knowledge (DOK) levels, as they each provide opportunities for test-takers to apply different cognitive strategies and demonstrate proficiency with social studies content. Each item type on the Social Studies Test may be presented either as a discrete item or as part of an item scenario in which two or three items pertain to a single stimulus.

Stimulus materials may include brief text, maps, graphs, tables, or other graphic representations of data or scientific concepts. Many of the brief texts featured in both discrete items and item scenarios will be drawn from texts reflecting "the Great American Conversation." These texts may be directly excerpted from founding documents, such as The Bill of Rights, or they may contain analyses of these documents. They may also be drawn from other more contemporary primary and secondary source documents (e.g., political speeches and commentary) that convey important concepts about American civics.

Multiple choice (MC) items are used to assess aspects of virtually every Social Studies Practice and Content Topic listed in the GED® Social Studies Assessment Targets. This item type continues to be a reliable, standardized method for measuring skills and knowledge at a range of cognitive levels. Unlike the multiple choice items on the 2002 Series GED® Test, the MC items on the new assessment only have four answer options rather than five.

Fill-in-the-blank (FIB) items on the Social Studies Test give test-takers the opportunity to construct a very brief response, like a single word or a short phrase, when potential answers have little variability. For example, this item type can be used when an item requires a test-taker to identify a particular data point on a chart reflecting economic trends. It can also be used to excerpt a word or phrase from a text to demonstrate understanding of an idea or vocabulary term that could be inferred from a brief textual stimulus.

Drop-down items with drop-down menu functionality embedded within a brief text are used to give test-takers opportunities to choose the correct response to complete statements. As with editing tasks in the RLA Test, test-takers are given the advantage of seeing the complete statements they create in an interactive manner onscreen. These items can measure many of the same skills that fill-in-the-blank items can, though they provide a selection of possible responses from which test-takers can choose. This item type is especially effective for assessing how well a test-taker can identify a logical conclusion drawn from text-based evidence or even make a generalization based on an author's argument.

Drag-and-drop items are another type of interactive task that require test-takers to move small images, words, or numerical expressions to designated drop targets on a computer screen. They may be used to assess how well a test-taker can make comparisons between concepts or representations of data or how well they classify or order information. For example, an individual drag-and-drop item may require a test-taker to place labels on a map to indicate important commodities produced in various regions. Other items might provide the test-taker an opportunity to place data points or labels drawn from a brief text onto a graph or chart.

Hot spot items consist of a graphic image with virtual "sensors" placed strategically within the image. They can be used to measure a test-taker's understanding of relationships between data points cited from a textual or graphic stimulus. They are also particularly effective for measuring a test- taker's ability to understand geographic concepts with regard to mapping. Other applications of hot-spot functionality might include asking test-takers to select data or points in a graph, chart, or table that support or refute a given conclusion stated in a brief textual stimulus.

The Extended response (ER) item on the Social Studies Test is a 25-minute task that requires test-takers to analyze one or more source texts in order to produce a writing sample. These ERs are scored on three dimensions, as outlined in the Extended Response Multi-trait Scoring Rubric. The first trait on the rubric pertains to how well test-takers analyze arguments and gather evidence from the source text in support of the positions that they take in their writing samples. The second trait scores the writing samples on the basis of how well the writing is developed and organized. The writing samples are also scored for how well test-takers demonstrate fluency with conventions of Edited American English, per the third trait on the rubric. On the Social Studies Test, the first trait of the rubric is scored on a three-point scale, and the second and third traits are each scored on a two-point scale. The prompts for the ERs are developed to elicit analytic writing that effectively uses evidence from the source text(s).

Assessment Targets for All Content Areas

Reasoning Through Language Arts (RLA) Assessment Targets

In alignment with career- and college-readiness standards, the GED® RLA assessment focuses on three essential groupings of skills:

- The ability to read closely
- The ability to write clearly
- The ability to edit and understand the use of standard written English in context

Because the strongest predictor of career and college readiness is the ability to read and comprehend complex texts, especially nonfiction, the RLA Test includes texts from both academic and workplace contexts. These texts reflect a range of complexity levels, in terms of ideas, syntax, and style. The writing tasks, or Extended Response (ER) items, require test-takers to analyze given source texts and use evidence drawn from the text(s) to support their answers.

Given these priorities, the GED® RLA Test adheres to the following parameters:

1. Seventy-five percent of the texts in the exam are informational texts (including nonfiction drawn from science and social studies as well as a range of texts from workplace contexts); 25 percent are literature.
2. The texts included in the test cover a range of text complexity, including texts at the career- and college-readiness level.
3. For texts in which comprehension hinges on vocabulary, the focus is on understanding words that appear frequently in texts from a wide variety of disciplines and, by their definition, are not unique to a particular discipline.
4. U.S. founding documents and “the Great American Conversation” that followed are required texts for study and assessment.
5. The length of the texts included in the reading comprehension component of the test vary between 450 and 900 words.
6. Roughly 80 percent of the items are written to a Depth of Knowledge cognitive complexity level 2 or 3; DOK level 4 is beyond the scope of the GED® test.
7. Reading and writing standards, such as those found in career- and college-readiness standards, are also measured in the GED® Social Studies Test, and the reading standards are measured in the GED® Science Test.

Reading Comprehension on the GED® RLA Test

The reading comprehension component of the GED® RLA Test measures two overarching reading standards that appear in the Common Core State Standards as Anchor Reading Standards 1 and 10, respectively:

- Determine the details of what is explicitly stated and make logical inferences or valid claims that square with textual evidence
- Read and respond to questions from a range of texts that are from the upper levels of complexity, including texts at the career- and college-ready level of text complexity

These two high-level standards broadly govern all aspects of passage selection and item development in the reading comprehension component of the GED® RLA Test. As candidates are asked to determine the main idea, the point of view, the meaning of words and phrases, and other inferences and claims, they are asked to do so based on texts that span a range of complexity, including texts at the career- and college-readiness level. The specific assessment targets that define the domain of the reading component of the GED® RLA Test and the connection to career- and college- readiness standards are described next.

The targets and indicators in the following tables are derived from Career- and College Readiness Reading Comprehension and Language Standards and govern the skills assessed in individual items.

Writing on the GED® RLA Test

The writing component integrates reading and writing into meaningful tasks that require candidates to support their written analysis with evidence drawn from a given source text(s) of appropriate complexity provided in the test. Also, given the growing demand and use of technology in all levels of post-secondary education and careers, the GED® test is administered by computer. Therefore, as in the reading component of the RLA Test, the following two high-level standards, which correspond with Common Core Anchor Standards 9 and 6, respectively, broadly govern all aspects of the writing tasks.

1. Draw relevant and sufficient evidence from a literary or information text to support analysis and reflection.
2. Use technology to produce writing, demonstrating sufficient command of keyboarding skills.

Candidate responses are scored by a multi-trait rubric that focuses on three elements:

- Trait 1: Analysis of Arguments and Use of Evidence
- Trait 2: Development of Ideas and Structure
- Trait 3: Clarity and Command of Standard English Conventions

The specific assessment targets that define the domain of the writing component of the GED® RLA Test and the connection to the CCR standards are described next.

Language Conventions and Usage on the GED® RLA Test

The language component of the GED® RLA Test measures a candidate’s ability to demonstrate command of a foundational set of conventions of standard English that have been identified as most important for career and college readiness by higher education instructors of post-secondary entry-level, credit-bearing composition courses. This core set of skills includes essential components of grammar, usage, capitalization and punctuation.

The GED® RLA Test includes editing items in an authentic context in which highlighted words or phrases appear in drop-down menus offering alternatives, which will include a clear best choice alongside common errors or misconceptions.

Mathematical Reasoning Assessment Targets

The GED® Mathematical Reasoning Test focuses on two major content areas: quantitative problem solving and algebraic problem solving. Evidence that was used to inform the development of the career- and college-readiness standards shows that instructors of entry-level college mathematics value master of fundamentals over a broad, shallow coverage of topics. National remediation data are consistent with this perspective, suggesting that students with a shallow grasp of a wide range of topics are not as well prepared to succeed in post-secondary education and are more likely to need remediation in mathematics compared to those students who have a deeper understanding of more fundamental mathematical topics. Therefore, the GED® Mathematical Reasoning Test focuses on the fundamentals of mathematics in these two areas, striking a balance of deeper conceptual understanding, procedural skill and fluency, and the ability to apply these fundamentals in realistic situations. A variety of item types is used in the test, including multiple choice, drag-and-drop, hot spot, and fill-in-the-blank.

The career- and college-readiness standards include Standards for Mathematical Practice, which describe the types of practices, or behaviors, in mathematics that are essential to the mastery of mathematical content. These standards form the basis of the GED® mathematical practice standards, which assess important mathematical proficiencies, including modeling, constructing and critiquing reasoning, and procedural fluency.

Given these priorities, the GED® Mathematical Reasoning Test adheres to the following parameters:

1. Approximately 45 percent of the content in the test focuses on quantitative problem solving, and approximately 55 percent focuses on algebraic problem solving.
2. The test includes items that test procedural skill and fluency as well as problem solving.
3. The contexts within which problem-solving skills are measured are taken from both academic and workforce contexts.
4. Approximately 50 percent of the items are written to a Depth of Knowledge cognitive complexity level of 2.
5. Approximately 30 percent of the items are aligned to a Mathematical Practice standard in addition to a content indicator.
6. The statistics and data interpretation standards are also measured in the GED® Social Studies and Science tests.
7. Candidates are provided with an on-screen calculator, the Texas Instruments TI-30XS Multiview scientific calculator, for use on most of the items on the 2014 GED® Mathematics Test. (The on-screen calculator is also provided for selected items on the Science and Social Studies tests.) Additionally, test-takers may bring their own personal handheld Texas Instruments TI-30XS scientific calculator.

Mathematical Practices

In addition to the content-based indicators, the GED® mathematics test will also focus on reasoning skills, as embodied by the GED® Mathematical Practices. The mathematical practices framework is based upon two sets of standards: the Standards for Mathematical Practice found in CCR standards for mathematics; and the Process Standards found in the Principles and Standards for School Mathematics, published by the National Council of Teachers of Mathematics.

The content indicators and mathematical practices found in the GED® Mathematical Reasoning Assessment Targets, though related, each cover different aspects of item content considerations. The content indicators focus on mathematical content, as typically seen in state standards frameworks and, to some extent, the career- and college-readiness standards for mathematics. The indicators describe very specific skills and abilities of which test-takers are expected to demonstrate mastery. In contrast, the mathematical practices focus more on mathematical reasoning skills and modes of thinking mathematically. Most of these skills are not content-specific, meaning that a mathematical practice indicator could be applied to items that cover a range of content domains (e.g., algebra, data analysis, number sense). The measurement of these skills is very much in keeping with the Career- and College-Readiness Standards for Mathematical Practice, which were created in order to “describe varieties of expertise that mathematics educators at all levels should seek to develop in their students.” The mathematical practices provide specifications for assessing real-world problem-solving skills in a mathematical context rather than requiring students only to memorize, recognize, and apply a long list of mathematical algorithms.

While we consider it crucial to assess both content and reasoning, it would be unrealistic to assert that each individual item could address both types of skills. To be sure, there are interrelated concepts to be found in the content indicators and the mathematical practices, especially in the areas of modeling and fluency, but not every item assessing a content indicator interacts seamlessly with a mathematical practice. Rather than force alignments, we seek to create items in which content and practice mesh well together. These items would primarily assess practice, with content serving as the context in which the practice is applied. Items of this type reflect the reasoning and problem-solving skills that are so critical to college and career readiness. Where this type of natural overlap between practice and content is not possible, other items assess the content indicators directly, thereby ensuring coverage of the full range of mathematical content on each test form.

Science Assessment Targets

The GED® Science Test focuses on the fundamentals of science reasoning, striking a balance of deeper conceptual understanding, procedural skill and fluency, and the ability to apply these fundamentals in realistic situations. In order to stay true to this intention, each item on the Science Test is aligned to one science practice and one content topic.

The science practices can be described as skills that are key to scientific reasoning in both textual and quantitative contexts. The science practices are derived from important skills enumerated in the career- and college-readiness standards as well as in The National Research Council’s Framework for K-12 Science Education.

The Science Test also focuses on three major content domains: life science, physical science, and Earth and space science. The science content topics, which are drawn from these three domains, provide context for measuring a test-taker's abilities to apply the reasoning skills described in the practices. The content topics focus on science that reflects both what is taught in many high school–level science courses and what is most relevant and useful for an adult population. To measure this content at a range of levels of complexity, several different item types are used in the test, including multiple choice, short answer, drag-and-drop, hot spot, and fill-in-the-blank.

Given these priorities, the GED® Science Test adheres to the following parameters:

1. Approximately 40 percent of the test focuses on life science, roughly 40 percent focuses on physical science, and approximately 20 percent focuses on Earth and space science.
2. The test includes items that test textual analysis and understanding, data representation and inference skills, and problem solving with science content.
3. Each item on the Science Test is aligned to both one science practice and one content topic.
4. Each item is also aligned to one Depth of Knowledge (DOK) level of cognitive complexity, based on the appropriate alignment to a science practice.
5. Approximately 80 percent of the items are written to a DOK level of 2 or 3; DOK level 4 is beyond the scope of the GED® test.
6. Problem-solving skills are measured in both academic and workforce contexts.
7. Approximately 50 percent of the items are presented in item scenarios, in which a single stimulus (which may be textual, graphic, or a combination of both) serves to inform two to three items. The rest of the items are discrete (i.e., stand-alone) items.

2014 GED® Test Science Content Topics

The science content topics describe key concepts that are widely taught in a variety of high school–level courses and are relevant to the lives of GED® test-takers. The content topics are designed to provide context for measuring the skills defined in the science practices section of this document.

As in the previous version of the GED® Science Assessment Targets, the science practices maintain a close relationship with the science content topics. More specifically, the primary focus of the GED Science Test continues to be the measurement of essential reasoning skills applied in scientific context. However, test-takers should still be broadly and generally familiar with each of the basic concepts enumerated in the science content topics and subtopics, and they should be able to recognize and understand, in context, each of the terms listed there. Nevertheless, test-takers are not expected to have an in-depth and comprehensive knowledge of each subtopic. Rather, the stimuli about which each question pertains will provide necessary details about scientific figures, formulas, and other key principles. For example, a question may include answer options and stimuli that contain specific terms drawn from the content subtopics; however, test-takers will never be asked to formulate their own definition of a term without the item providing sufficient contextual support for such a task.

Focusing Themes

As previously noted, the two science focusing themes [(a) Human Health and Living Systems and (b) Energy and Related Systems] have been selected to ensure that the test covers a wide range of important scientific topics, but they are also intended to function like a lens by drawing focus to a distinct subset of ideas within each content topic. That is, items from any of the three content domains of life science, physical science, and Earth and space science can pertain to one of these two themes, but content that falls outside the spheres of these themes will not appear on the Science Test.

Social Studies Assessment Targets

The GED® Social Studies Test focuses on the fundamentals of social studies reasoning, striking a balance of deeper conceptual understanding, procedural skill and fluency, and the ability to apply these fundamentals in realistic situations. In order to stay true to this intention, each item on the Social Studies Test is aligned to one social studies practice and one content topic.

The social studies practices can be described as skills that are key to scientific reasoning in both textual and quantitative contexts. The practices come from important skills specified in career- and college-readiness standards, as well as in National Standards for History.

The Social Studies Test also focuses on four major content domains: civics and government, United States history, economics, and geography and the world. The social studies content topics, which are drawn from these four domains, provide context for measuring a test-taker's ability to apply the reasoning skills described in the practices. The content topics focus on key concepts that reflect both what is taught in many high school–level social sciences courses and what is most relevant and useful for an adult population.

To measure this content at a range of levels of complexity, several different item types are used in the test, including multiple choice, drag-and-drop, hot spot, and fill-in-the-blank. Additionally, the Social Studies Test will feature one extended- response task that requires test-takers to analyze arguments and use evidence found within brief excerpts from primary and secondary source texts.

Given these priorities, the GED® Social Studies Test adheres to the following parameters:

1. Approximately 50 percent focuses on civics and government, 20 percent focuses on United States history, 15 percent focuses on economics, and 15 percent focuses on geography and the world.
2. The test includes items that assess textual analysis and understanding, data representation and inference skills, and problem solving using social studies content.
3. Social Studies Test items align to one social studies practice and one content topic.
4. Each item aligns to one DOK level, based on appropriate alignment to social studies practice.
5. Approximately 80 percent of the test items are written to DOK level 2 or 3; DOK level 4 is beyond the scope of the GED® test.
6. Problem-solving skills are measured in both academic and workplace contexts.

7. Approximately 50 percent of the test items are based on scenarios in which a single stimulus (textual, graphic, or a combination of both) serves to inform two or three items; the remaining approximately 50 percent of the items are discrete (i.e., stand-alone) items.

Social Studies Content Topics

The social studies content topics describe key concepts that are widely taught in a variety of high school–level social studies courses and are relevant to the lives of GED® test-takers. They focus, in particular, on American civics and government. They are designed to provide context for measuring the skills defined in the social studies practices section of this document.

The social studies practices maintain a close relationship with the social studies content topics. More specifically, the primary focus of the GED® Social Studies Test continues to be the measurement of essential reasoning skills applied in a social studies context. However, test-takers should still be broadly and generally familiar with each of the basic concepts enumerated in the social studies content topics and subtopics, and they should be able to recognize and understand, in context, each of the terms listed there. Nevertheless, test-takers are not expected to have an in-depth and comprehensive knowledge of each subtopic. Instead, the stimuli about which each question pertains will provide necessary details about social studies-related figures, events, processes, and concepts. For example, a question may include answer options and stimuli that contain specific terms drawn from the content subtopics; however, test-takers will never be asked to formulate their own definition of a term without the item providing sufficient contextual support for such a task.

Focusing Themes

As previously noted, the content topics for the Social Studies test focus on two main themes [(a) Development of Modern Liberties and Democracy and (b) Dynamic Responses in Societal Systems], each applied across the four domains in the social studies arena (i.e., civics and government, U.S. history, economics, and geography and the world). These themes have been selected to ensure that the test covers a wide range of important concepts and ideas in social studies, but they are also intended to function like a lens to draw focus to a distinct subset of ideas within each content topic. Content that falls outside the parameters of these themes is not included in the new Social Studies test.

Reporting Category Descriptions for Content Areas

One of our goals for the score reports on the new GED® assessment is to provide additional information about areas of strength and developmental need. In order to generate this information, we have grouped indicators from each content area’s assessment targets into reporting categories. Points that test-takers earn in each category contribute to sub-scores from each content area. The primary purpose of these sub-scores is to give guidance to both test-takers and their instructors so that each test-taker can successfully achieve his or her GED® test credential.

Below are a series of brief descriptions of the types of skills that are assessed in each individual reporting category. The descriptions of each reporting category are illustrative, not exhaustive.

Reasoning Through Language Arts Reporting Category Descriptions

Reporting Category 1: Analyzing and creating text features and technique (and example skills)

- Analyzing essential elements of both literary and informational texts
- Analyzing how parts of a text fit into the overall structure
- Analyzing an author's point of view or purpose and the rhetorical techniques authors use to advance meaning

Reporting Category 2: Using evidence to understand, analyze, and create arguments (and example skills)

- Analyzing arguments and using evidence to demonstrate close-reading skills
- Identifying and evaluating an author's underlying premise(s)
- Distinguishing between supported and unsupported claims, and assessing the validity of an author's reasoning
- Comparing two or more sources to analyze differences in use of evidence, interpretation, overall impact, and other modes of argument making
- Writing their own arguments in which they synthesize details, draw conclusions, and apply information from given source texts

Reporting Category 3: Applying knowledge of English-language conventions and usage (and example skills)

- Demonstrating sufficient command of essential standard English conventions and usage Correcting common grammatical or usage errors
- Demonstrating fluency with these skills in their own writing

Mathematical Reasoning Reporting Category Descriptions

Reporting Category 1: Quantitative problems in rational numbers (and example skills)

- Demonstrating fluency with operations using rational numbers
- Using rational numbers to formulate solutions to problems set within real-world contexts
- Solving problems with rational numbers that involve proportionality

Reporting Category 2: Quantitative problems in measurement (and example skills)

- Engaging with geometric figures in a variety of graphic presentations
- Engaging with descriptive statistics in a variety of graphic presentations
- Using formulas or decomposition to calculate perimeter, area, surface area, and volume of figures
- Using descriptive statistics to summarize and compare data sets and understand concepts relating to basic theoretical probability

Reporting Category 3: Linear equations and expressions (and example skills)

- Writing linear mathematical expressions and equations that correspond to given situations
- Evaluating the expressions for specific values of the variable
- Solving linear equations, inequalities, and systems of linear equations and finding the equation of a line with varying criteria
- Interpreting slope of a line as rate of change or unit rate

Reporting Category 4: Function concepts and nonlinear expressions and equations (and example skills)

- Understanding and applying the concept of a function
- Using function notation
- Translating a variety of representations of a function, including tables and equations
- Solving quadratic equations
- Interpreting key features of both linear and nonlinear functions

Science Reporting Category Descriptions

Reporting Category 1: Analyzing scientific and technical arguments, evidence, and text-based information (and example skills)

- Analyzing scientific and technical texts to determine hypotheses, data, and conclusions
- Citing evidence within the text that supports the hypotheses and conclusions
- Thinking critically about texts to determine facts versus opinion and evaluate an author’s claims
- Summarizing scientific texts and evaluating key terms and relationships among concepts within the text

Reporting Category 2: Applying scientific processes and procedural concepts (and example skills)

- Applying scientific reasoning skills to a broad range of content
- Creating and explaining the features of a hypothesis
- Conducting and critiquing experimental procedures
- Making generalizations and drawing valid conclusions from experimental data

Reporting Category 3: Reasoning quantitatively and interpreting data in scientific contexts (and example skills)

- Using mathematical techniques to interpret and analyze scientific data
- Engaging with data displayed in various graphic formats
- Making calculations using a variety of statistical and probability techniques
- Identifying proper measurement practices and units, including conversions between units

The Science reporting categories are organized according to the Science Practices rather than the Science content indicators. This method of organization has been chosen because the Science Practices are integrated into every item on the Science test and represent thinking and reasoning skills that are critical for adults to master. Although the content Topics and Subtopics are also reflected in all items, the Science content areas are too numerous for the test to be able to provide reliable and meaningful reporting data on them. Test-takers, however, will be receiving much more detailed information than ever before on both the skills they possess and those they need to develop. With this additional information, adult educators will be in a position to focus their work with test-takers on critical skill development needs.

The reporting information provided by the 2014 GED® test is one of the most important elements of the new assessment system. Gaining a firm understanding of the reporting categories on the GED® test will help adult educators in planning how they can best help adult learners to gain the skills they will need to be successful on the test and in the future pathway they ultimately pursue.

Social Studies Reporting Category Descriptions

Reporting Category 1: Analyzing and creating text features in a social studies context (and example skills)

- Analyzing primary and secondary sources for various purposes
- Identifying aspects of a historical document that reveal the author's point of view or purpose
- Distinguishing between unsupported claims and those that are grounded in evidence necessary for understanding concepts in the social sciences
- Determining the meaning of domain-specific words used in context

Reporting Category 2: Applying social studies concepts to analysis and construction of arguments (and example skills)

- Applying social-studies-specific reasoning skills to a variety of tasks
- Examining the relationships among people, environments, events, processes, and ideas and accurately describing the chronological and/or causal nature of the relationships
- Comparing different ideas within social studies disciplines such as civics and economics, and examining the implications of these ideas
- Producing writing that thoroughly and logically develops an idea, claim, or argument based on primary and/or secondary source texts
- Supporting contentions with specific textual evidence from the source texts and demonstrating an understanding of the contexts in which these documents were written

Reporting Category 3: Reasoning quantitatively and interpreting data in social studies contexts (and example skills)

- Analyzing data presented in a wide variety of formats, including maps, graphic organizers, photographs, and political cartoons
- Integrating analyses of quantitative data with analyses of written information to inform their understanding of the topic at hand
- Accurately using and interpreting graphs in order to analyze the differing ways in which variables are related to one another

The Social Studies reporting categories are organized according to the Social Studies Practices rather than the Social Studies content indicators. This method of organization has been chosen because the Social Studies Practices are integrated into every item on the Social Studies test. While the content indicators are also reflected in all items, the Social Studies content Topics and Subtopics are too numerous for the test to be able to provide reliable and meaningful reporting data. Test-takers, however, will be receiving much more detailed information than ever before on both the skills they possess and those they need to develop. With this additional information, adult educators will be in a position to focus their work with test-takers on critical skill development needs.

The reporting information provided by the 2014 GED® test is one of the most important elements of the new assessment system. Gaining a firm understanding of the reporting categories on the GED® test will help adult educators in planning how they can best help adult learners to gain the skills they will need to be successful on the test and in the future pathway they ultimately pursue.

Test Development and Construction Process

Content for the GED testing program is developed in “waves.” An item development plan is designed for each wave of development. This plan drives the number of items and passages that are developed. Every item used on the GED test undergoes a rigorous review process. The review process includes many levels of review by Pearson School, GED Testing Service, Content and Fairness Committees, and Key Reviews. After the item is developed, it is rendered in XML format. Additional reviews take place to ensure the accuracy and quality of the items. Once items have passed numerous levels of quality checks, they are ready to be field-tested.

Field-testing rules are established by psychometricians to ensure the statistical integrity of each item. After field testing, research scientists evaluate the statistical data and conduct a data review. Items are evaluated during data review by research scientists and assessment specialists. Statistical data are evaluated to ensure that the content of each item is fair to all demographic groups and meets statistical and quality standards. Items are then available for forms construction. After test forms construction, the items are reviewed again in a computer-based test environment. The forms undergo numerous quality checks that include content, functionality, and scoring reviews. Forms that surpass all quality checks are published on the Pearson VUE platform and made available to test-takers.

Below is a high-level summary of the item development/test construction process.

A. **Develop Content Validity Guide**

B. **Design Item Development Plan**

C. **Complete Item/Passage Writer Training**

D. **Complete Pre-Committee Reviews** (steps 2-8 repeat until items/passages are approved)

    1. Content Triage Review and feedback to Item/Passage Writers
    2. Content Review and editing by Assessment Specialist
    3. Copy Edit review
    4. Universal Design review by Alternate Assessment Specialists
    5. Fact Checking review by Research Librarians
    6. Concept Art Creation by Artists
    7. Content Review and editing by Assessment Specialist
    8. Content Review by Assessment Specialist

E. **Complete Content and Fairness Committee Reviews**

F. **Complete Post-Committee Reviews** (steps 2-7 repeat until items/passages are approved)

    1. Post-committee reconciliation of edits
    2. Application of edits by Assessment Specialist
    3. Copy Edit review
    4. Fact Checking review by Research Librarians
    5. Concept Art Editing by Artists
    6. Content Review and editing by Assessment Specialist
    7. Content Review by Assessment Specialist

G. **Complete Item Rendering**

    1. XML item rendering by XML Designer
    2. Art Creation by Artists
    3. Compiling of items by XML Designer

H. **Complete Rendered Item Review** (steps 1-7 repeat until items/passages are approved)

    1. Content Review by Assessment Specialist
    2. Copy Edit review
    3. Application of edits by XML Designer
    4. Compiling of items by XML Designer
    5. Content Review by Assessment Specialist
    6. Copy Edit review
    7. Content Review by Assessment Specialist

I. **Complete Test Construction of Field Test Forms** (steps 1-10 repeat until forms are approved)

    1. Pull and approve test forms
    2. Complete and approve test maps
    3. Complete assessment XML

4. Content Forms Review by Assessment Specialist
5. Forms review by Copy Edit
6. Application of edits by XML Designer
7. Compiling of Forms by XML Designer
8. Forms Review by Assessment Specialist
9. Forms review by Copy Edit
10. Forms Review by Assessment Specialist
11. Key Reviews
12. Quality testing and review
13. Forms Publishing

J. **Complete Test Construction of Operational and Readiness Forms** (steps 1-10 repeat until forms are approved)

1. Pull and approve test forms
2. Complete and approve test maps
3. Complete assessment XML
4. Content Forms Review by Assessment Specialist
5. Forms review by Copy Edit
6. Application of edits by XML Designer
7. Compiling of Forms by XML Designer
8. Forms Review by Assessment Specialist
9. Forms review by Copy Edit
10. Forms Review by Assessment Specialist
11. Key Reviews
12. Quality testing and review
13. Forms Publishing

Item Field-Testing

The new GED® test was launched in January 2014 with three operational forms. In order to create new sets of operational and GED Ready® forms, GED Testing Service implemented an embedded field test design. The embedded field testing began in October 2014 for the Mathematical Reasoning, Science, and Social Studies tests. Field testing for Reasoning Through Language Arts began in January 2015. Items are randomly selected from a pre-specified pool of field test items and embedded randomly within the operational tests. All field test items are randomly selected from a larger pool, which remains active until sufficient sample sizes are obtained. Further details are described below.

Mathematical Reasoning

The Mathematical Reasoning (Math) test has two sections, one in which a calculator is not allowed, followed by a section that allows a calculator. The Math field test pool rotates between pools of non-calculator items (embedded within the non-calculator section of the test) and calculator items (embedded within the calculator section). Each pool is further subdivided into sets. One set comprises non-scenario-based items while the other contains scenario-based items, technology-enhanced items, and non-scenario-based items to balance the number of items in each set. One item from each set is embedded within each Math test.

Science & Social Studies

Both the Science and Social Studies tests utilize a rotating pool of field test items.[8] A single field-test item is randomly selected from the pool and randomly embedded within the operational tests. Items that are based on a scenario set are embedded as a single item. Field-test items do not appear within an operational scenario set.

Reasoning Through Language Arts

The Reasoning Through Language Arts test utilizes a rotating pool of field test items.[9] Each pool is further subdivided into two sets. One set comprises passage-based items (four items per passage), and another set comprises cloze items (five items). One group of items is selected randomly from one of the two sets. The field-test items can appear in the section before or after the Extended Response item. The items associated with the passage are always presented in a fixed order.

Calibration and Scaling of Field Test Items

All field test items are scaled using item response theory. During the first scaling procedure (for each content area) in 2014, the operational item parameters were also rescaled to align more closely with the operational testing population. For each content area, all operational and field test items within the first field test pool were freely calibrated. The operational items were then used to find scaling constants to link the current scale to the Standardization and Norming Study (SNS) base scale. The scaling transformation was applied to all freely calibrated operational and field-test items. The item bank was updated with these new operational and field-test item parameter estimates. For subsequent field-test analyses by content area, all field-test items within a content area pool are calibrated simultaneously with the operational items fixed to their updated operational item parameter estimates.

[8] Excludes the Social Studies Extended Response Item.

[9] Excludes the Reasoning Through Language Arts Extended Response Item.

Chapter 4: Standardization and Norming, Scaling, and Equating

This chapter describes the processes of norming, scaling, and equating the GED® test. **Standardization** and **norming** refer to the process of administering the GED® test to a national sample of recent high school graduates (the norm group) to establish typical scores for that norm group. **Scaling** refers to the process of transforming raw GED® test scores (e.g., number of items correctly answered) to scaled scores that possess desirable qualities useful for comparing scores on the same content area tests across different forms. **Equating** refers to the statistical process of adjusting test scores so that the level of performance indicated by a particular scaled score is consistent from form to form. These three processes are described more thoroughly in this chapter.

Prior to standardization, norming, and equating, we provide a brief review of the field-testing methodology that preceded the launch of the 2014 GED® test. The 2012 Stand-Alone Field Test (SAFT) was an extensive effort to obtain feedback on the performance of newly developed test items. Further details of the field test can be found in the 2012 Stand-Alone Field Test: Technical Manual (GED Testing Service, 2013).

2012 Stand-Alone Field Test

GED Testing Service performed a large-scale, SAFT operation in 2012. The SAFT was a key step in developing the operational assessment. It provided an opportunity to assess the quality and performance of newly developed items from both content and statistical/psychometric perspectives. The results helped determine whether many of the assumptions about items and scoring were accurate and where improvements in item development or scoring were necessary. Data from the SAFT were used to assess item-level aspects, such as key checks, scoring accuracy, item difficulty, item-total correlations, and item fit (according to the item response theory model). Data from extended response (ER) and short answer (SA) items were also used to train and calibrate the automated scoring engine.

GED Testing Service recruited a large sample of adults to participate in the SAFT. The participants represented various regions across the United States.[10] The field test included multiple waves of administration across several months. Items were assembled into fixed modules (modules were not necessarily representative of an operational form). A spiral plan was implemented that allowed a minimum of 250 responses to each item type. Technology-enhanced items and ER and SA items were overlapped across multiple modules in order to obtain larger sample sizes.

[10] Demographic characteristics of the SAFT participants can be found in the 2012 Stand-Alone Field Test: Technical Manual (GED Testing Service, 2013).

Item Quality Checks

One of the main purposes of the SAFT was to gauge the performance of newly developed test items. The SAFT was the first opportunity to gain any type of performance data from these new items. The quality of the items was reviewed thoroughly such that any items not performing to pre-determined standards would be excluded from future test forms. Additionally, data were aggregated to determine if certain item types were not performing well. Content specifications were also reviewed for gaps in item coverage or content areas that were proving difficult to measure with current items or item types.

SAFT items fell into two broad scoring groups: machine-scored and human-scored. Machine-scored items included all those scored automatically by Pearson VUE's test delivery system. Human-scored items included the ER and SA items. A number of analyses were performed on the machine-scored data. Item analyses included the calculation of the item difficulties (p-values) and item-total correlations (point-biserial), as well as the response distribution frequencies. Items were also calibrated using the Master's Partial Credit Model, and therefore Rasch item difficulties (RID) and fit statistics (infit and outfit) were reviewed. The following criteria were used to flag machine-scored items for review:

- p-value
- point-biserial correlation
- distribution of responses
- percent omitted
- percent omitted given sequence to gauge speededness

The following statistics were calculated on human-scored items:

- item mean
- point-biserial
- score distribution
- percent omitted/condition codes

Response times per item type and content area were also calculated and assessed. This information was used to help determine the amount of testing time required for the operational test. Additional information was collected from the SAFT participants after completing the study. Specifically, participants were asked about their experiences, including their level of academic preparedness for the test content; their computer familiarity; the clarity of tutorial, calculator, and other test instructions; and the time allotted to complete the test modules.

Key Outcomes of the Stand-Alone Field Test

The SAFT yielded several different types of beneficial information. Most notable was feedback on item quality and the scoring processes. Adjustments were also made to test timing based on the response times demonstrated during the study. These issues are discussed more fully in the following sections.

Item Content and Statistical Review

First, GED Testing Service was able to assess the quality of the items, the size of the eligible item pool, and the utility of different item formats. Items were analyzed for content and statistical characteristics and summarized to gauge the quality of the overall item pool from which initial operational forms were developed. Additionally, the utility of various item types, including several technology-enhanced formats, was assessed.

Extended Response and Short Answer Scoring

GED Testing Service utilizes an automated scoring procedure for both ER and SA items. Since the automated scoring occurs outside the test delivery system, an extensive technological infrastructure was created to route these items to and from the scoring engine, all in a very short time. A system for human scoring was also integrated into the procedure so that scores could be presented to the automated scoring engine for calibration. Operationally, human scoring will continue to be necessary on a limited basis. The SAFT provided an opportunity to test many of these systems, including the flow of data from location to location.

Changes to Scoring Rubric and Answer Scoring Guides

Performance data from the SAFT resulted in some changes in the Extended Response Scoring Rubrics for the Reasoning Through Language Arts test. Specifically, GED Testing Service decided to remove the highest score for all three traits. Similarly, the Social Studies rubric was amended to remove the highest score point on Trait 1. Additionally, GED Testing Service opted to shift the entire scale down a point so that each trait in RLA is scored on a 0-2 scale (rather than 1-3) and, in Social Studies, Trait 1 is scored on a 0-2 scale and Traits 2 and 3 are scored on a 0-1 scale (rather than 1-2). This shift allows for a more accurate point allocation, because responses that receive a score of 0 on any given trait are defined as insufficient for that trait. Slight changes to the language within the scoring rubrics were also made to increase clarity for the human scorers and parallelism across score points. However, the essence of the criteria by which each test-taker's response was evaluated did not change.

Adjusted Scoring Keys

A primary outcome of the SAFT was the adjustment to scoring keys. Multiple-choice items were reviewed to determine whether the response distributions were aligned with expectations (i.e., higher-ability test-takers could identify the correct response). Fill-in-the-blank items were thoroughly reviewed to ensure all variations of a correct response were included in the scoring key. In some cases, items were dropped entirely if it was determined that the response variations were too varied or complex to adequately score.

2013 Standardization and Norming Study

Scores obtained from the GED® test are used to make claims about how the candidate performed relative to the population of graduating high school seniors. Therefore, the passing standard is based on normative information obtained from such a population. GED Testing Service conducted the 2013 Standardization and Norming Study (SNS) to fulfill this purpose.

Data for the SNS were collected from a national sample of recently graduated high school students. The methodology used for the SNS data collection was unlike any used in previous GED® test norming studies. Previous studies involved sampling high schools and subsequently administering the test to students within the schools. Given that previous GED® test administrations were paper-based, this methodology generally worked well.

GED Testing Service has recently migrated to a primarily computer-based administration. The 2014 GED® test was developed with this strategy in mind. Therefore, the SNS needed to be administered similarly. GED Testing Service conducted some preliminary analyses and determined that performing a computer-based administration within high schools would have been extremely difficult.

In 2012, GED Testing Service performed a stand-alone field test in which a for-hire model was used to recruit participants (GED Testing Service, 2013). GED Testing Service decided to utilize a similar methodology for the 2013 SNS. The for-hire methodology is described in more detail next.

For-Hire Model and Recruiting

GED Testing Service used a for-hire model to obtain a national sample of graduating high school students. Instead of sampling high schools and then students within high schools, GED Testing Service utilized a sampling plan to recruit high school students across the nation and then paid students for participating in the SNS.

Compensation

Participants received various levels of compensation depending on their level of involvement and persistence in testing. Those participants who completed their original spiral assignment received $200, plus a completion bonus of $100. Participants who did not complete their original spiral assignment received $14 for each full-length test and $7 for each GED Ready® test and the tutorial.

GED Testing Service wanted to encourage participation early in the study so that participation could be monitored. If participation was too low, it could have jeopardized the study results. Therefore, GED Testing Service offered a $50 bonus to participants who scheduled their first appointment by June 30, 2013, and took their first two modules by July 15. In addition, a $100 Pearson book voucher was provided to all participants who completed their assigned modules by July 20, 2013.

Participants could earn additional compensation by taking additional test modules. This incentive was provided only to those participants who completed their originally assigned modules by July 31, 2013. Those qualified participants could earn additional compensation by taking up to10 additional modules. Each participant was compensated according to the number of additional modules taken. The compensation for additional testing modules was $20 for Forms A, B, and C of the Mathematical Reasoning, Science, and Social Studies tests; $25 for Forms A, B, and C of the Reasoning Through Language Arts tests; and $10 for Forms RA and RB for all content areas. The maximum compensation for the additional testing was $145.

As an additional incentive to encourage participants to complete their assigned tests and to encourage current testers to invite their friends and former classmates to participate in the study, GED Testing Service offered a $25 referral bonus for eligible test-takers who referred new test-takers who completed the study. (The referral bonus was provided over and above any additional compensation that the test-taker might have earned in the study.) To be eligible, both the "referring test-taker" and the "referred test-taker" must have completed all of their individually assigned testing modules in their testing schedules.

As an additional incentive to improve participant effort, awards were given to the highest score on the test battery (comprising Form A) as well as each content area test (Form A only). Awards included an Apple® iPad® Wi-Fi only, 16 GB mobile digital device for the battery-level awards, and an Apple® iPad® Mini Wi-Fi only, 16 GB mobile digital device for the individual test awards. Test-takers were eligible for only one prize and received the highest-level prize that they were qualified to win.

As an additional incentive to encourage participants to complete their assigned tests, test-takers who completed all of their assigned test modules in their testing schedule by the end of the study window received an unofficial score report with the scaled score, percentile rank, and performance level for the test forms that comprise the GED® test battery.

Finally, high schools were asked to encourage students to participate in the SNS. A $250 award was given to the top 10 high schools with the most recruits.

Sample

A two-stage, complex sampling plan was created for the study. The first sampling phase involved sampling counties across the United States from which the high school seniors would be recruited. Once a large pool of high school seniors was developed, a final probability-based sample of students was obtained.

County Sample

The first task was to determine where to recruit participants for the study. Partly on the basis of the final sample size requirements, the contractor recommended that a stratified probability sample (proportional to size, based on the total number of high school graduates) of U.S. counties should be selected first. Eight strata were created by crossing four Census regions with metro/non-metro status.

A probability sample of 100 counties was drawn in late 2012. The preference was for counties that contained at least one Pearson VUE testing center site. The county sampling frame was based on the restricted-use 2007-2008 and 2008-2009 National Center for Education Statistics' Common Core of Data Dropout and Completer files for public schools and the restricted-use 2009–2010 Private School Survey files for private schools.[11]

Student Sample

Once the sample of counties was selected, high school students were recruited for the study (see Recruiting of Participants section below). From the pool of applicants, a sample of 1,013 current high-school graduates was drawn for the SNS. The sampling methodology assumed that the graduates were a pseudo-random sample from all graduates within the United States, i.e., that each graduate had a propensity of self-recruiting into this graduate frame, and this propensity was entirely determined by their gender, race/ethnicity, urbanicity, and achievement levels.

Recruiting of Participants

High school students within the selected sample of counties were recruited for the SNS. (At the time of the recruitment process, the participants were still enrolled in high school. However, the participants were recent high school graduates at the time the SNS was conducted.) GED Testing Service hired an outside contractor to develop and implement an online application for the SNS. A website was created specifically for the SNS, which contained links to the study application, frequently asked questions, details about the study, and modes of contacting the study administrators for questions or concerns. Links to a Facebook page and Twitter were also provided.

The contractor also completed phone calls to the high schools within the sampled counties. During the calls, the contractor explained the nature of the study and requested the school promote the opportunity to participate to students via job boards, flyers, emails, parent-teacher associations, ads within school newspapers, or posting links to the study website.

Students interested in participating in the SNS were directed to an online application. The application requested basic demographic information such as full names, home address, and other contact information. Responses to a number of other questions were obtained, such as likelihood of graduating within the study window, academic background, high school information, plans after graduation, and parents' education levels. All of these variables were used to screen potential students for participation and for the final sampling weights.

[11] Access to the restricted-use frames was granted to the contractor under a special license from the National Center of Education Statistics.

Weighting Procedures

Several types of weights were generated for differing parts of the instrument:

- Four weights for each of the four components of the instrument
- A 'battery weight' for those graduates who completed all four components of the instrument.

The adjustment for each of the five sets of weights was to the following control cells for which control totals are available:

- 1—g=1,…,4 being the four Census regions;
- 2—u=1,…8 being urbanicity (u=1—city, large; u=2—city, midsize; u=3—city, small; u=4—suburb, large city; u=5—suburb, midsize city; u=6—suburb, small city; u=7—town; u=8—rural);
- 3—i=1,2 being gender (i=1—male; i=2—female);
- 4—j=1,..,5 being race/ethnicity (j=1—Hispanic; j=2— non-Hispanic Black, j=3—non-Hispanic Asian or Pacific Islander, j=4—non-Hispanic American Indian, j=5—non-Hispanic White or multiple race or other);
- 5—k_1=1,..,9 being 9 subgroups for combined Scholastic Aptitude Test (SAT)[12] or composite ACT[13] score (k_1=1—no SAT and no ACT taken; k_1=2—0th to 12.5th percentile combined SAT or composite ACT, k_1=2—12.5th to 25th percentile combined SAT or composite ACT, …., k_1=9—87.5th percentile to 100th percentile combined SAT or composite ACT)[14];

An initial weight $W_{cs}^{(init)}$ was set equal to the 3,361,915 (the estimated total number of high-school graduates in spring 2013), divided by the sample size for each exam component. The summation of these weights was then 3,361,915. Suppose the summation of these weights is taken within each marginal cell defined by the categories above. These are model-unbiased estimates of the total numbers of graduates within these cells (assuming the graduate self-selection process is effectively random: all other things being equal, each graduate has an equal propensity of being on the frame):

$$\hat{X}^{(1)}(g) = \sum_{cs \in S^{(1)}(g)} W_{cs}^{(init)}, \quad g = 1, \ldots.,4 \qquad \hat{X}^{(2)}(u) = \sum_{cs \in S^{(2)}(u)} W_{cs}^{(init)}, \qquad u = 1, \ldots.,8$$

[12] This is the summation of the scores for SAT Critical Reading and Mathematics test scores.

[13] This is a composite of ACT English and Mathematics test scores.

[14] Graduates who took both the SAT and ACT and whose percentile categories were not consistent were dealt with as follows. A mean value was taken of the SAT percentile category and the ACT percentile category. If this mean value was an integer, this became the joint category. If this mean value was a fraction, a random number was drawn to assign to the higher or lower category. For example, if the SAT category is 2 and the ACT category is 4, then the joint category was 3. If the SAT category was 2 and the ACT category was 5, then with one-half probability the recruit was assigned to joint category 3 and with one-half probability the recruit was assigned to joint category 4.

$$\hat{X}^{(3)}(i) = \sum_{cs \in S^{(3)}(i)} W_{cs}^{(init)},\ \ i = 1{,}2 \qquad\qquad \hat{X}^{(4)}(j) = \sum_{cs \in S^{(4)}(j)} W_{cs}^{(init)},\ \ j = 1, \dots, 5$$

$$\hat{X}^{(5)}(k_1) = \sum_{cs \in S^{(5)}(k_1)} W_{cs}^{(init)},\ \ k_1 = 1, \dots, 9$$

The S sets (e.g., $S^{(1)}(g)$) contain all of the graduates at that level for that dimension. The weights $W_{cs}^{(init)}$ were calibrated to sum up to the true values of these control totals$X^{(1)}(g), g = 1, \dots., 4,\ X^{(2)}(u),\ p = 1, \dots., 8, \dots\dots\dots.. X^{(5)}(k_2),\ k_2 = 1, \dots, 9.$

The weights were raked in such a way so that the final weights $W_{cs}^{(f)}$ were as close as possible to the initial weights while satisfying the control equations:

$$X^{(1)}(g) = \sum_{cs \in S^{(1)}(g)} W_{cs}^{(f)},\ \ g = 1, \dots., 4 \qquad\qquad X^{(2)}(u) = \sum_{cs \in S^{(2)}(u)} W_{cs}^{(f)},\ \ p = 1, \dots., 8$$

$$X^{(3)}(i) = \sum_{cs \in S^{(3)}(i)} W_{cs}^{(f)},\ \ i = 1{,}2 \qquad\qquad X^{(4)}(j) = \sum_{cs \in S^{(4)}(j)} W_{cs}^{(f)},\ \ j = 1, \dots, 5$$

$$X^{(5)}(k_1) = \sum_{cs \in S^{(5)}(k_1)} W_{cs}^{(f)},\ \ k_1 = 1, \dots, 9$$

The inverse of these final weights $p_{cs}^{(f)} = \left\{W_{cs}^{(f)}\right\}^{-1}$ was the estimated propensity of a graduate self-recruiting onto the frame, conditional on their Census Region, urbanicity, gender, race/ethnicity, and achievement levels as measured by SAT and ACT tests.

The raking process was iterated with a trimming process that restrained the minimum and maximum weights[15]. Each trimming process was followed by a raking process, followed by another trimming process, until convergence was achieved.[16] This process provided weights that satisfied the control totals and also satisfied lower and upper bounds on the weights. The trimming process introduced a potential for bias, but this was controlled by the raking process, which guaranteed the weights to be in conformity with the control totals. This trimming process

[15] The full set of raking results are not included here, but they are included in the full Sampling report.

[16] The algorithm was stopped when the trimming process was changing no more than 0.001 of the largest weight within each trimming cell, or when 200 iterations were done.

was done separately within each of the achievement-level (based on SAT and ACT scores) cells.

Any weights that were smaller than the median weight within each trimming cell divided by X_1=3.5 were trimmed, as well as any that were larger than the median weight times X_2=3.5, unless this resulted in more than five percent of the graduate weights being trimmed on either side. If the latter situation occurred, the cutoff for X_1 or X_2 was set so that exactly five percent of the graduate weights in the cell were trimmed back on either side. The trimmed weights were the median weight divided by X_1 on the low side, and the median weight multiplied by X_2 on the high side. This process reduced the variability from extreme weights, while the bias was limited since no more than 10 percent of the units were trimmed.

Initial base weights were equal to the population divided by the sample size. These simple base weights were put through the raking process to achieve initially raked weights $W_{cs}^{(a)}(1)$. These initial raked weights were trimmed to get$w_{cs}^{(a)}(2)$, and subsequently raked to get $W_{cs}^{(a)}(2)$. This process continued until convergence. The final weights $W_{cs}^{(a)}$ satisfied completely the control totals, and satisfied the trimming constraints within certain limits.

A necessary condition was that all observations had non-missing values for the benchmark characteristics Census Region, urbanicity, gender, race/ethnicity, and SAT/ACT scores. Missing values were imputed where necessary. In particular, for the SAT/ACT scores, there were two sets of scores, one set of scores given by the applicants at the time that they took the exam, and another set of scores given by the applicants at the time that they originally applied. If one set of scores was present and the other was missing, the non-missing scores were used. If both sets of scores were present (for SAT or ACT), the most recent of the two scores was taken.

If the SAT/ACT category was missing (not one or the other set of scores present for each the application or exam period, and no positive affirmation that the recruit did not take either test by putting a zero score), SAT/ACT category was imputed. This process was done by hot-deck imputation, using grade point average (GPA) categories within public/private school status, poverty category, and urbanicity. GPA category was used as a primary predictor of SAT/ACT category, allowing for systematic differences in GPA distributions between public and private schools, and schools by poverty category and urbanicity.

Test Administration (Spiral) and Scheduling

In order to establish the flexibility necessary to support a for-hire model and to provide some randomization and counter-balancing, a spiraling plan was created. Examinees were randomly assigned to one of six spirals and one of four groups that specified the order in which the tests were administered. Total testing time ranged from 16.5 hours to 17.18 hours, depending on the spiral. Some participants were assigned to a spiral with 12 test modules, while some received 13 tests modules. However, examinees were given the option to take additional exams outside of their assigned spiral.

The spiral was designed such that all participants received a core set of forms (one from each content area) and a random subset of the remaining forms, and the two GED® Ready forms. In addition, each participant received a brief survey as one of the modules.

Testing occurred under standardized conditions within authorized Pearson VUE Testing Centers. Test-takers were able to schedule their tests at their convenience during the testing window (July 1-Aug 10, 2013), in the order in which the tests were assigned. The allotted testing times for the SNS were the same as those for the operational test, and the administration of the SNS was very similar to that of the operational test in 2014. However, we do acknowledge some differences, such as taking GED Ready® forms after operational forms.

Scoring

All multiple-choice, fill-in-the-blank, editing passages (RLA only), and 1- and 2-point technology-enhanced items were scored by the test delivery platform. Extended response (ER) and short answer (SA) items were either 100 percent scored by Pearson Scoring Center (PSC) or scored using a PSC/automated-scoring (AS) combination. For this combination, the top end of the score distribution was sent to PSC while the rest of the distribution was automatically scored. In addition, responses deemed as outliers (those responses that AS could not confidently score) were sent to the PSC for scoring.

The ER and SA scoring during the SNS was otherwise similar to the operational scoring procedures. Each SA item is worth 3 points and has its own unique scoring rubric developed specifically for that item. Examinees were assigned a score of 0-3 or, when applicable, a condition code. All examinees who received a condition code were assigned a score of zero for that item. A list of all possible condition codes is as follows:

- CT: Response exclusively contains text copied from source text(s) or prompt
- OT: Response shows no evidence that test-taker has read the prompt or is off-topic
- IN: Response is incomprehensible
- NE: Response is not in English
- BL: Response has not been attempted (blank, etc.)

The RLA ER item was scored using an analytic rubric with three traits, in which each trait was assigned a score of 0-2, resulting in range of 0-6 points per RLA ER item. The Social Studies ER item was scored with a three-trait rubric where the first trait was assigned a score of 0-2 and the second and third traits were assigned a score of 0-1, resulting in a range of 0-4 points per Social Studies ER item. (It should be noted that blanks were automatically scored by the test delivery system and were not sent for human scoring.)

Monitoring Effort

Several indicators of test-taker effort were used throughout the study. Data files were provided daily such that analyses were be performed routinely. The effort monitoring strategy involved the calculation of several indicators, which were then reviewed by GED Testing Service psychometricians. These analyses are described next.

Rapid-Guessing Analysis

Item response times were used as a measure of the amount of time each test-taker spent on each item. Rapid-guessing (Schnipke, 1995, 1996; Schnipke & Scrams, 1997) is a term used to describe behaviors in which a test-taker responds to an item so quickly that it is unlikely he or she was able to read and comprehend the item content (i.e., the test-taker guessed the response). The following steps were implemented to flag rapid guessing at the item level, as well as to summarize the test-taker’s performance across all items within a test module.

1. Flag any response time less than 3 seconds.
2. For each person by form, calculate the proportion of items not flagged for rapid-guessing. This is the response time effort (RTE) score.
3. Flag if RTE score is less than 55 percent.

Timing Out and Score Percentage

Another effort indicator was to see whether an individual used the full amount of testing time while exhibiting very low performance. This indicator was similar to the rapid-guessing indicator, except that the test-taker spent more than 5 seconds on the items. This indicator was a bit less reliable at flagging low effort, since very low ability test-takers could easily exhibit the same behaviors. Therefore, test-takers flagged for this indicator underwent additional scrutiny. The following process was implemented.

1. Identify individuals who scored approximately chance-level (based on number of multiple-select items).
2. Identify individuals whose total module time was within 2 minutes of the allotted time.
3. Individuals flagged for both criteria above were examined at the item level.
    a. Item-level response times were examined for obvious issues, such as spending the allotted time on just one item.
    b. Reponses to open-ended items were examined to determine whether the test-taker made effort.
    c. Individuals who were flagged multiple times across different modules were noted.

Omitted Items

Test-takers may have omitted an item response for a number of reasons. For example, some test-takers simply may not have had the requisite skills or knowledge to provide a reasonable response and therefore decided to skip the item entirely. It is also possible that some test-takers simply did not try on some items and skipped over them. Therefore, we used the following process to flag individuals for omitted items.

1. The number of omitted items was summed for each individual, by module.
2. Individuals were flagged if the percentage of omitted items exceeded 50percent.

Word Counts and Response Times on Extended Response Items

We suspected that some responses to ER items would suffer from lower effort levels, given that the ER requires a different cognitive load and response process. We discussed the issue with content experts and gathered some expectations for a minimum-level response. The following process was used to flag individual performances on the ER items.

1. Any ER consisting of fewer than 100 words was flagged.
2. Response times associated with the ER items were also examined. For example, if the respondent used the full amount of allotted time and only responded with a few words, then the individual was flagged for that module.

The four indicators described earlier (rapid-guessing analysis, timing out and score percentage, omitted items, and word counts and response times on ER items) were combined into a single data set for review. Test-takers who were flagged for multiple indicators were examined carefully by the psychometrics staff. We looked at each flag and the reasons why the test-taker was flagged and considered alternative explanations for the behaviors. In addition, we reviewed site-level reports uploaded by testing proctors. These reports occasionally provided observational evidence that a test-taker was not trying—e.g., sleeping at the work station or clearly not paying attention.

As mentioned, these indicators were reviewed daily since GED Testing Service received a nightly data upload from Pearson VUE. Decisions regarding whether to remove individuals from further testing were made by the next day, and Pearson Human Resources promptly notified individuals that they were no longer eligible to participate. Removed individuals were noted (by ID number) and excluded from further analyses. Our approach was very conservative, in our opinion, and resulted in very few individuals being removed from the study and analyses. After review of the flagged records, many did not show conclusive evidence of low effort. For example, many ER responses of less than 100 words were, in fact, on topic; few were not. After several weeks, the effort monitoring was reduced to a weekly basis.

Calibration

All items administered during the Standardization and Norming Study were placed on a common underlying scale using item response theory (IRT). All examinees were required to take a core set of content area test forms and a randomized subset of the remaining tests forms. The target sample size for the core form was 1,500. The remaining forms were calibrated using a sample of 750 (500 for the GED Ready® forms).

All forms within a content area were calibrated simultaneously using an incomplete data matrix. This concurrent calibration assumed randomly equivalent groups (established via the random assignment to spirals) and resulted in all forms being on a common scale. Master's Partial Credit Model was used to support IRT item parameter estimation for all GED® content area tests. Under this model, the probability of a person *n* obtaining a score *x* on a given *m*-step item *i* is estimated using the following equation:

$$P_{xni} = \frac{exp\sum_{j=0}^{x}(\theta_n - \delta_{ij})}{\sum_{k=0}^{m_i} exp\sum_{j=0}^{x}(\theta_n - \delta_{ij})}$$

Where *x* = 0,1, …., *m* and δ_{ij} = the *j*th step difficulty for polytomous item *i*. For dichotomous items this reduces to

$$P_{ni} = \frac{exp(\theta_n - D_i)}{1 + exp(\theta_n - D_i)}$$

given that

$$\sum_{j=0}^{0}(\theta_n - \delta_{ij}) \equiv 0$$

WINSTEPS 3.74 was used for all IRT analyses. A traditional item analysis was performed on all multiple choice items just prior to the calibration phase. The purpose of the item analysis at this point was to identify any potential administration or scoring issues. Item p-values, item-total correlations, score distributions, and percent omitted were reviewed for each item. At the same time, all technology-enhanced items underwent an adjudication process in which the frequency distribution of correct and incorrect responses was reviewed. Content specialists then reviewed the distributions to ensure the scoring was correct and that items were functioning as expected. Item means, item-total correlations, score distributions, and percent omitted/condition codes were also reviewed.

Because examinees took multiple forms and because there is item overlap across operational forms and on the two GED Ready® forms, examinees were occasionally exposed to the same item more than once. For calibration purposes, an examinee's first attempt at an item was included; subsequent attempts were excluded.

Scaling

After the calibration phase was completed, the item parameters from the concurrent calibration were set as fixed, and the core set of forms were scored using IRT (Rasch and Master's Partial Credit Model). For each content area, the raw-score-to-theta score conversion was established. Next, the raw-score-to-scale-score table was established using the core set of forms as the base form. The theta score passing standard and Honors-level cut score were used to determine the linear scaling transformations by specifying two scale score points. The scale score of 150 was mapped to the theta score passing standard cut (found through empirically based methods by GED Testing Service), and the scale score of 170 was mapped to the theta score with Honors cut-score (found through content-based methods by GED Testing Service). The following equation established the linear transformation:

$$SS_\theta = \frac{170 - 150}{\theta_H - \theta_P}\theta + \left[150 - \frac{170 - 150}{\theta_H - \theta_P}\theta_P\right]$$

where θ_H is the theta score associated with Honors cut-score and θ_P is the theta score associated with the passing-standard cut-score. The resulting scaling transformation was used to convert thetas into scaled scores as follows:

$$SS_\theta = a\theta + b$$

Lowest and highest obtainable scaled score values were set to 100 and 200, respectively. The scaling transformations were subsequently applied to each form.

Chapter 5: Establishing the Performance Standard on the 2014 GED® Test

The 2014 GED® test has three performance levels, and the passing standard for high school equivalency has been normed and standardized using a national sample of high school graduates from the class of 2013. The GED® test established two *benchmarks* (also known as standards, or cut scores) for each of the four content areas. The minimum benchmark, referred to in this Technical Manual as the *Passing Standard,* indicates the minimum level of performance necessary to meet the requirements for a high school–level credential as demonstrated by the empirical performance of high school seniors who have recently graduated. As previously noted, throughout the history of the GED® test, the cut score for the Passing Standard has been defined as one that is “not so high as to hold adult learners to a higher standard than that of graduating high school seniors nor so low as to threaten the validity and credibility of the GED® credential."

Attainment of the Passing Standard in all content areas is intended to result in the awarding of a high school equivalency credential. The higher of the two performance benchmarks is *GED® with Honors.* This second performance level represents knowledge and skills that are indicative of successful performance outcomes in first-year credit-bearing courses in post-secondary education programs.

The 2014 GED® test supports three performance levels:

- **Performance Level 1** is associated with scores below the GED® test Passing Standard (100 to 149 scaled score points).
- **Performance Level 2** is associated with scores at or above the Passing Standard (150 to 169 scaled score points).
- **Performance Level 3,** the highest level, is associated with performance indicative of career- and college-readiness known as the GED® with Honors (170 scaled score points and above).

Performance Level Descriptors: Tools for Understanding Performance

Along with determining the performance standards, performance level descriptors (PLDs) have also been developed to describe the knowledge and skills represented by each of the three performance levels (Below Passing, Passing Standard, and GED® with Honors [representing career- and college-readiness]) for each content area. These PLDs flesh out the meaning of high school equivalency and help test-takers identify the skills that they possess in each content area consistent with career- and college-readiness as well as identify those skills they must attain for improved performance. The content- and skills-based information and its presentation on the GED® Enhanced Score Report is designed to help adults and their instructors plan for the acquisition of advanced skills through the GED® testing process. PLDs can be found on the educator resource page at http://www.gedtestingservice.com/2014testresources.

Establishment of Performance Levels on the GED® Test

The GED® Passing Standard is the point on the score scale that represents a reasonable and sufficient level of performance expected of adult learners on the targeted academic knowledge, skills, and abilities (KSAs), given the performance of a national sample of recent high school graduates.

Historically, the passing standard on the test has been set at various levels, based on the contemporary best judgment of experts. These levels were set to account for the fact that when high school students took the GED® test as part of a norming process the test did not carry the same high stakes as it did for adult learners actually seeking their credentials. Therefore, the current passing standard for adult learners needs to take into account the impact that different levels of motivation have on performance of the high school graduates who participated in the SNS demonstrate.

When the first GED® test was established in the 1940s, the passing standard was set at the point where 80 percent of high school graduating seniors would have passed and 20 percent would have failed the test. For the 2002 Series, almost 60 years later, this level was set (through the judgment of experts and stakeholders) at the point where 60 percent of high school graduating seniors would have passed and 40 percent would have failed, based on the circumstances at the time and the performance characteristics of the particular sample of students in the study.

GED Testing Service considered several factors in our plan when estimating the GED® Passing Standard for each content area sub-test. The following list shows the criteria into considered for determining the recommended cut scores.

1. Historical passing standard for the content area tests (e.g., 18th percentile) and for the battery (e.g., 40th percentile)
2. Historical pass rates (e.g., immediately after releasing a new edition and over time during the tenure of an edition)
3. Observed motivation and effort statistics regarding participation in the GED® Test Standardization and Norming Study
4. Estimates from content specialists, policy specialists, and others related to the percentage of graduating high school seniors they would expect to pass the test, given factors such as motivation, degree of instruction and/or preparation, administrative conditions, and computer proficiency
5. Average GED® test performance associated with recent high school graduates' self-reported content area grades in their senior year
6. Average GED® test performance associated with various grade point average (GPA) values
7. Expected pass rates at the battery level given different estimated pass rates at each content area

The GED® Passing Standard

The *GED® Passing Standard* refers to the minimum level of performance in each content area that must be achieved in order for a GED® test-taker to be eligible to earn a GED® test credential. Currently, this standard is set to 150 for each content area test.

The Passing Standard for each subject area test for the 2014 GED® test was determined empirically. These empirical cuts were determined by examining the performance of test-takers who had a self-reported GPA of 2.99 or below (i.e., below a B average). Our rationale for this decision is that we wanted the passing score for the new GED® test to be reasonable and reflective of performance typical of graduating high school seniors. Using a higher GPA would most likely have resulted in our setting unrealistic expectations for adult learners, and using a lower GPA would most likely represent the student performances well below average and therefore would not make a sound point of comparison for adult learners seeking a high school equivalency credential. Hence, we feel confident that basing the passing standard on those students who achieved a GPA of B – and lower sets a point of comparison that is both reasonable and attainable for the GED® test-taking population, and yet will still result in a credential that is meaningful to end users, such as employers and college admissions.

We also conducted a content-based standard-setting approach with GED Testing Service's expert content specialists. This process was designed to create additional validation evidence to compare against the empirically based cuts. We implemented a "bookmark" procedure that aimed to connect the judgment-based task of setting cut scores to the measurement model and to connect test content with performance level descriptors (Mitzel, Lewis, Patz, & Green, 2001). Using a difficulty index, the items are ordered from the easiest to most difficult in an item booklet. The ordered item booklet shows one item per page. The purpose of the ordered item booklets, as stated by Lewis et al. (1998), is "to help participants foster an integrated conceptualization of what the test measures, as well as to serve as a vehicle to make cut score judgments" (p. 7). The cut scores based on this bookmark item-mapping approach proceeded in several rounds. Once the final set of cut scores was determined, GED Testing Service lowered the cut scores by one standard error to take measurement error into account.

The empirically based cut scores were compared against the final content-based cut scores. A review of the results allowed us to determine that the empirical passing standard cuts were corroborated with qualitative evidence derived from the test content. During a meeting with a GED® Policy Board, we shared the results of the empirically based results as compared to the content-based results. As displayed in Table 2 below, the cut scores for Science and Social Studies resulted in the same raw score cut for both the empirical approach and the content-based approach. For Mathematical Reasoning and Reasoning Through Language Arts, the empirical approach resulted in a cut -score that was one raw score point higher than the content-based approach.

Table 2. Raw Score Associated with the GED® Passing Standard by Content Area for Content-Based and Empirically Based Methods

Content Area	Empirically Based Cuts	Content-Based Cuts
Mathematical Reasoning	21	20
Science	18	18
Social Studies	18	18
Reasoning Through Language Arts	26	25

GED® with Honors Benchmark

The generation of a benchmark that provides test-takers with information regarding their "readiness" for careers and college is a new component of the GED® test. At the time of this writing, GED Testing Service has not labeled this benchmark as a "career- and college-readiness standard." This is because additional validity evidence is needed in order to support such an inference. In the meantime, GED Testing Service has referred to this benchmark as GED® with Honors. The current benchmark for GED® with Honors is set at 170 for each content area test.

As with the GED® Passing Standard, it is important to note the caveats associated with the GED® with Honors benchmark. GED Testing Service does not currently state that performance relative to the cut-score provides for any predictive inferences regarding readiness for careers or college. Collecting evidence to support this claim will take time, which explains why the college-and career-readiness indicator will be a future enhancement feature.

GED Testing Service also acknowledges and understands that the operational definition of "readiness" is fluid over time. The national conversation regarding what constitutes readiness for careers and college will continue to evolve as both research and policy also continue to evolve. In that sense, changes to the benchmark may be necessary in future years.

Results

A detailed set of results associated with the cut scores for the Passing Standard and the GED® with Honors cut scores were discussed during the October 2013 Stakeholder Meetings. We shared information regarding the reporting scale, the percentage of points required for each performance level, the impact data for each performance level, percentile ranks, and other summary information that helped illustrate how the SNS sample test results compared to other indicators of interest. Performance level descriptors and items that represented the KSAs just below and just above the cut-score were also displayed and discussed in order to help further define the definition of the cut score.

Table 3Table 3 shows summary data from the standard-setting process. As shown in Table 3 each content area requires a scaled score of 150 to achieve the Passing Standard and a scaled score of 170 to achieve the GED® with Honors level. The test summary section shows the percentage of points that must be earned to score at the Passing Standard and GED® with Honors performance levels. As an example, for Mathematical Reasoning, test-takers must earn 43 percent of the points to earn the Passing Standard or 78 percent of the points to score at the GED® with Honors level.

The section titled Impact Data shows the percentage of test-takers from the weighted sample in the SNS that scored at each performance level. For example, for Mathematical Reasoning, 28 percent of test-takers were in the first performance level (Below Passing), 53 percent of test-takers scored in the second performance level, and 19 percent of test-takers scored in the highest performance level (GED® with Honors).

The next section of Table 3 represents the percentile ranks. The percentile rank ranges from 1 to 99. Percentile rank shows the percentage of graduating high school seniors who participated in the SNS and earned this score or lower. As an example, the percentile rank for the Passing Standard on the Mathematical Reasoning test is 31, and the percentile rank for GED® with Honors is 80.

The next section of the table shows the average score on the SAT of the test-takers in each performance level in the SNS. The SAT has Quantitative and Verbal sub-tests that can be used as a comparison with the GED® sub-tests for Mathematical Reasoning and Reasoning Through Language Arts. The average SAT Quantitative score of test-takers who scored in Performance Level 1 on the GED® Mathematical Reasoning test was 399. The average SAT score for test-takers in Performance Level 2 was 490, and the average score of test-takers in Performance Level 3 was 642. Similar SAT Verbal summaries are presented in the column for Reasoning Through Language Arts. The final section of the table contains similar information for the ACT exam.

Table 3. Summary Data Illustrating the Passing Standard and the GED® with Honors Cut Scores

	Mathematical Reasoning	Science	Social Studies	Reasoning Through Language Arts
Cut-scores on the Scaled Score Metric				
Passing Standard	150	150	150	150
GED with Honors	170	170	170	170
Test Summary				
% of points required to achieve the GED® Passing Standard	43%	45%	41%	40%
% of points required to achieve GED® with Honors	78%	75%	73%	74%
Impact Data				
Level 1: Below Passing	28%	27%	30%	31%
Level 2: Passing Standard	53%	61%	51%	58%
Level 3: GED with Honors	19%	13%	19%	11%
Percentile Ranks				
Passing Standard	31	28	31	29
GED with Honors	80	86	81	87
SAT Summaries				
Average Score	SAT Math	NA	NA	SAT Reading
Level 1: Below Passing	399	NA	NA	419
Level 2: Passing Standard	490	NA	NA	517
Level 3: GED with Honors	642	NA	NA	652
ACT Summaries				
Average Score	ACT Math	ACT Science	ACT Reading	ACT Reading
Level 1: Below Passing	15	15	17	15
Level 2: Passing Standard	21	22	22	23
Level 3: GED with Honors	29	30	29	30

Next Steps

In the years following the initial identification of these performance levels for each content area, GED Testing Service will continue to work with post-secondary institutions, including those providing workforce certification programs, to validate them. Through a program of research that will follow test-takers through a range of post-secondary pathways, data that will provide empirical evidence of the relationship between performance on the GED® test and performance in credit-bearing post-secondary courses will be collected.

Chapter 6: Scoring the GED® Test

Historically, GED Testing Service has used Classical Test Theory to build and score the GED® test. Early discussions surrounding the development of the 2014 GED® test included whether to use item response theory (IRT). Ultimately, it was decided that IRT could provide some benefits to the program. As described in Chapter 4, Master's Partial Credit Model was used to calibrate all field test and operational test items during the 2012 Stand-Alone Field Test and 2013 Standardization and Norming Study. GED Testing Service uses number-correct scoring, which is transformed to a scaled score using IRT.

In previous years, the scoring of the GED® test was decentralized. Each jurisdiction was responsible for contracting its paper-based test scoring with a GED Testing Service–approved scoring center. As the GED® test has migrated to a computer-based delivery system, the decentralized scoring process was no longer necessary. Instead, scoring is largely automated by the test delivery platform and other external, automated-scoring processes.

Scoring within the Computer-Based System

The majority of test items are scored within the computer-based delivery system, called Athena. Scoring rules are implemented within the item-level XML programming, which is carried out by the Athena system as test-takers respond to items. All multiple choice, fill-in-the-blank, hot spot, drag-and-drop, cloze, and drop-down items are scored directly by Athena. For certain items, the item scores are double-weighted (i.e., a score of 0 or 2). For editing passages (drop-down items), each item contains several responses. The responses are individually assessed as correct or incorrect, and a single score is reported for each editing passage.

Scoring of Extended Response and Short Answer Items

The extended response (ER) and short answer (SA) items are not scored by the Athena system. Instead, the responses are immediately routed to an external, automated-scoring system developed by Pearson Knowledge Technologies (KT). The only exception is when an ER or SA item is left completely blank by a test-taker, for which the response receives a score of zero for each trait (or total score for SA items) and is not routed to KT. In most cases, KT returns an ER or SA score within several hours, thus dramatically reducing the time a candidate must wait to receive a test score compared to previous GED® test series.[17]

In some cases, KT is unable to provide a score for an ER or SA response. In these cases, the responses are routed to Pearson Scoring Center for verification. These responses receive a condition code and are automatically assigned a zero score.

[17] ER and SA responses that require human scoring may take longer. In these cases, the candidate is notified that the test score is under additional review for quality purposes.

Condition codes result from the following:

- Response exclusively contains text copied from source text(s) or prompt
- Response shows no evidence that test-taker has read the prompt or is off-topic
- Response is incomprehensible
- Response is not in English
- Response has not been attempted (blank, pictures, etc.)

Scoring rubrics for the Reasoning Through Language Arts and Social Studies ER items are provided in Appendixes A and B, respectively.

Training the Automated-Scoring Engine

KT's automated essay-scoring technology is based on Latent Semantic Analysis (LSA), a machine-learning method that acquires and represents knowledge about the meaning of words and documents by analyzing large bodies of natural text. In addition to the underlying LSA technology, the KT engine includes additional customized development and proprietary mathematical techniques to optimize LSA for specific automated-scoring applications.

The KT algorithms require a training, or calibration, for each GED® test E) or SA) item. Calibration for the GED® test began with the 2012 Stand-Alone Field Test (SAFT). Responses to the ER and SA items were collected from the sample and evaluated by the KAT scoring engine. However, the SAFT did not elicit sufficient data for training the KT scoring engine, mainly due to a lack of responses earning the high score points. In addition, the SAFT sample was somewhat different from the operational, adult testing population.

The ER and SA items were again administered during the 2013 Standardization and Norming Study (SNS). During the SNS, items were either 100 percent scored by Pearson Scoring Center or scored using a PSC/KT combination. Item responses at the top end of the score distribution were sent to the PSC, while the rest of the distribution was scored by KT. In addition, responses deemed as outliers (those responses that KT could not confidently score) were sent to PSC for scoring. Data from the SNS were used to further calibrate the KT scoring engine.

Initial Analysis Period

As mentioned, the KT engine was calibrated using data from both the SAFT and SNS samples. Neither of these samples fully represented the variation seen in the adult testing population. Therefore, an initial audit period was established for the first several months of operational testing. During the initial audit period, additional ER and SA item responses were collected from the operational testing population. The ER and SA items were randomly assigned to each test-taker. The responses were scored initially by KT and subsequently routed to Pearson Scoring Center for human scoring (and resolution scoring, if necessary). The agreement rates were estimated and other analyses were performed to ensure the accuracy of the KT scoring system.

Score Reporting

GED Testing Service provides score reports online to anyone who takes a GED® test. The score reports provide the following information:

- Tells test-takers their scaled score, percentile rank, and what skills they did well on
- Identifies the specific skills the test-taker missed on the test
- Provides test-takers with targeted study recommendations synced to the specific skills missed and the full set of skills needed to improve a score

In addition, the score report provides one-click access to synced study materials available in print and online. Test-takers can browse and choose from study materials to customize their study recommendations.

Chapter 7: Reliability

Reliability refers to the consistency, or stability, of the test scores over replications. For example, if a given test yields widely discrepant scores for the same individual on separate test administrations, and the individual has not changed significantly on the measured attribute, then the scores on the test are not reliable. Conversely, if a test produces the same or similar scores for an individual on separate administrations, then the scores from the test are considered reliable. Reliability is inversely related to the amount of measurement error associated with test scores. That is, the more measurement error that exists in test scores, the less reliable the test scores.

Because reliability is a crucial index of test quality, test developers are required to evaluate and report the reliability of their test scores. Several procedures are used to evaluate reliability, and each accounts for different sources of measurement error and thus produces different reliability coefficients. One measure of reliability is *internal consistency,* which reflects the degree to which all the items on the test measure the same attribute or characteristic. Internal consistency coefficients are a convenient method of estimating reliability when parallel forms are not available. Another form of reliability is *decision consistency,* which reflects the consistency in which the test categorizes a test-taker into a scoring category. In this case, the GED® test has three classifications based on two benchmarks: the high school credential and GED® with Honors. Both internal consistency and classification consistency are discussed in more detail below.

Internal Consistency Reliability

Cronbach's coefficient alpha is a commonly used estimate of internal consistency because it can be applied in a number of measurement scenarios (e.g., various score scales and item formats).

The formula for the coefficient alpha reliability coefficient (α) is:

$$\alpha = \frac{k}{k-1}\left[1-\left(\frac{\Sigma\sigma_i^2}{\sigma_t^2}\right)\right] \qquad (7.1)$$

where *k* = the number of items on the test, σ_i^2 = the variance of item *i*, and σ_t^2 = the variance of the total scores on the test.

Coefficient alpha ranges from zero to one. As can be seen from Equation 7.1, three factors can affect the magnitude of the coefficient: the homogeneity of the test content (affects σ_i^2), the homogeneity of the examinee population tested (affects σ_t^2), and the number of items on the test (*k*). Tests comprising items that measure similar (i.e., homogenous) content areas will have higher reliability estimates than tests comprising items measuring diverse content areas because the covariance among the items is likely to be lower when the items measure widely different concepts or skills. Conversely, examinee populations that are highly homogenous can reduce the magnitude of the coefficient because the covariance among the items is limited by the amount of total variance in the examinee population.

The range of internal consistency reliability estimates is listed in Table 4.

Table 4. Internal Consistency Estimate Ranges, GED® Test 2014

Content Area	English
Mathematical Reasoning	.83 – .88
Reasoning Through Language Arts	.81 – .84
Science	.76 – .81
Social Studies	.75 – .80

Table 5. Internal Consistency Estimate Ranges, GED®, English, by Select Demographics, 2014

Demographic	Mathematical Reasoning	Reasoning Through Language Arts	Science	Social Studies
Male	.83 - .88	.80 - .84	.77 - .81	.76 - .81
Female	.82 - .87	.76 - .84	.74 - .80	.72 - .78
Non-White	.83 - .88	.80 - .84	.76 - .80	.74 - .78
White	.82 - .88	.79 - .82	.74 - .79	.73 - .80
Hispanic/Latino	.79 - .86	.76 - .83	.71 - .78	.70 - .77
Non-Hispanic/Latino	.83 - .88	.81 - .84	.77 - .81	.75 - .80

Standard Error of Measurement

The standard error of measurement (SEM) is an estimate of the average amount of error that is associated with scores derived from a test. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) defines the SEM as “the standard deviation of an individual’s observed scores from repeated administrations of a test (or parallel forms of a test) under identical conditions. Because such data generally cannot be collected, the [SEM] is usually estimated from group data” (pp. 223-224). The SEM is often used to describe how far an examinee’s observed test score may be, on average, from his or her “true” score (i.e., a score that is free from measurement error). Therefore, a smaller SEM is preferable to a larger one. The SEM can be used to form a confidence interval around an observed test score to suggest a score interval within which an examinee’s true score may fall. Because the SEM is the standard deviation of a hypothetical, normal distribution of measurement errors, in most cases, it is expected that an examinee’s observed score will be found within one SEM unit of his or her true score about 68 percent of the time.

The SEM is a function of the standard deviation of the test scores and of the reliability of the test scores. The equation for the SEM is:

$$SEM = \sigma_t \sqrt{1 - r_{tt}} \qquad (7.2)$$

where σ_t = the standard deviation of test scores, and r_{tt} = the reliability coefficient. From Equation 7.2, it can be seen that a test with a small standard deviation and large reliability yields a smaller SEM. Because the SEM is a function of the standard deviation of test scores, it is not an absolute measure of error; rather, it is expressed in raw score units. Therefore, unlike reliability coefficients, SEM cannot be compared across tests without considering the unit of measurement, range, and standard deviation of the tests' raw scores.

Estimates from 2014 indicate that the SEM ranges from 3 to 4 (scaled score metric) across all content area test forms.

Conditional Standard Errors of Measurement

As described above, the SEM provides an estimate of the average amount of error associated with an examinee's observed test score. However, the amount of error associated with test scores may differ at various points along the score scale. For this reason, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) state:

> When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (Standard 2.14, p. 46)

Conditional standard errors of measurement (CSEM) were generated using the WINSTEPS® software program. Each raw score point is associated with a CSEM. The scaling constants used to convert raw scores to scaled scores were also applied to the CSEM, thus allowing one to interpret the CSEM on the scaled score. Conditional standard errors of measurement are reported within plus/minus five scaled score points of the passing standard (150) and GED® with Honors (170) benchmark (see Table 6). The CSEM are consistent across all current operational forms.

Table 6. Selected Conditional Standard Errors of Measurement, GED® and GED® with Honors, by Content Area

Scale Score	Mathematical Reasoning	Reasoning Through Language Arts	Science	Social Studies
145	3	3	4	4
146	3	3	4	4
147	3	3	4	4
148	3	3	4	4
149	3	3	4	4
150	3	3	4	4
151	3	3	4	4
152	3	3	4	4
153	3	3	4	4
154	3	3	4	4
155	3	3	4	4
165	4	3	4	4
166	4	3	4	4
167	4	3	4	4
168	4	3	4	4
169	4	3	4	4
170	4	4	4	4
171	4	4	5	4
172	4	4	5	4
173	4	4	5	4
174	4	4	5	5
175	4	4	5	5

Reliability of Extended Response and Short Answer Scores

The extended response (ER) and short answer (SA) items are 100 percent scored by an automated-scoring (AS) engine.[18] GED Testing Service has performed a number of calibrations and analyses to ensure the reliability and accuracy of the AS engine. One of these procedures was an initial audit period in which all ER and SA responses were scored by both the AS engine and by human scorers. The section below describe the analyses that occurred during the initial audit period (January to March 2014).

Typically, human-human analyses (comparisons between the two human raters) are used as a point of reference for AS-human analyses (represents comparisons between automated-scoring and human-scoring scores). Human-human data were obtained from the 2012 Stand-Alone Field Test and the 2013 Standardization and Norming Study. AS-human scoring data were obtained from the initial audit period. Analyses began after approximately 500 valid human-scoring and AS scores were obtained and score distributions were deemed appropriate.

[18] Exceptions occur when the automated-scoring engine is unable to score a response. In these cases, the response is human scored.

The following analyses were conducted for each ER and SA trait, where R_1 is the first rater and R_2 is the second rater of the analyses. Human-human and AS-human scoring data were analyzed separately, followed by a comparison of the results.

1. Agreement rates:
    a. Exact agreement between two raters
    b. Adjacent agreement between two raters
    c. Non-agreement between two raters
2. Quadratic kappa: a comparison between the mean square error of rating pairs that are supposed to agree (X_1, Y_1) and those that are unrelated (X_1, Y_2), using the following equation

$$KAPPA = \frac{E([X_1 - Y_1]^2)}{E([X_1 - Y_2]^2)},$$

3. Standardized mean differences: the mean rating for rater 1 ($\bar{X}_{R1}$) and rater 2 ($\bar{X}_{R2}$), and their standard deviations sd_{R1} and sd_{R2}, respectively, are standardized using the following equation:

$$\bar{Z} = \frac{|\bar{X}_{R_1} - \bar{X}_{R_2}|}{\sqrt{\frac{sd^2_{R_1} + sd^2_{R_2}}{2}}},$$

4. Correlations: the ratio between the covariance between rater 1 and rater 2 [$cov(R_1, R_2)$] and the product of their standard deviation sd_{R1} and sd_{R2}, respectively, as shown in the following equation:

$$r_{R_1,R_2} = \frac{cov(R_1, R_2)}{sd_{R_1} * sd_{R_2}}$$

5. Two-way frequency tables

The results of these analyses were presented by each item-level trait within content area (excluding Mathematical Reasoning, which does not contain any ER or SA items). Evaluation of the analyses was conducted at the trait level because total score analyses could mask potential trait-level issues. Item-total score analyses were provided as secondary analyses to ensure nothing unusual occurred when item trait scores were summed.

The following criteria were used to assess the quality of the scoring.[19]

1. Agreement rates: Generally, AS-human scores should show agreement rates similar to those of human-human scores. However, agreement rates are biased by the number of score points of the scales, and standardized measures such as Kappa, standardized mean differences, and correlations are preferred.
2. Quadratic kappa (Kappa): Items were flagged if the AS-human scoring quadratic kappa was less than 0.7, and/or if the reduction in the quadratic kappa from human-human to AS-human was greater than 0.1.
3. Standardized mean differences: Items were flagged if the standardized mean difference was greater than .15.
4. Correlations: Items were flagged if the AS-human score correlation was less than 0.7 and/or if the reduction in the correlation from human-human to AS-human was greater than 0.1.
5. Two-way frequency tables: Items were flagged if the human-AS frequency tables differed substantially from the human-human frequency tables. The direction of the differences was noted.

Decision Consistency

Another form of reliability is decision consistency, which reflects the consistency with which the test categorizes a test-taker into a scoring category. In this case, the GED® test has three classifications based on two benchmarks: the high school credential and GED® with Honors. Standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) states:

> When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test-takers who would be classified in the same way on two replications of the procedure. (p. 46)

Decision Consistency Based on Content Area Tests

The GED® test forms include both dichotomously and polytomously scored test items. Therefore, decision consistency can be estimated using the Livingston and Lewis (1995) procedure. Classification accuracy is estimated by comparing the observed score distribution with a true score distribution predicted by the model. Similarly, a comparison between the observed score distribution and an alternate form predicted by the model provides a decision consistency estimate. The Livingston and Lewis procedure is implemented using the BB-Class software program developed by Brennan (2004).

Classification reliability can be estimated for each benchmark individually or concurrently. The GED® with Honors benchmark is considered a lower-stakes standard compared to the GED® Passing Standard. Therefore, to date, GED Testing Service has only examined the classification reliability at the test passing standard level. Ranges of classification accuracy and consistency estimates are presented in Table 7 and Table 8, respectively, for all English- and Spanish-language test forms.

[19] The criteria were obtained from Williamson, Xi, and Breyer (2012) and Ramineni and Williamson (2013).

Table 7. Classification Accuracy in GED® English- and Spanish-Language Test Forms

Content Area	Proportion of Correct Classifications	False Positive Rate	False Negative Rate
Mathematical Reasoning	.93 - .99	<.01 - .06	<.01 - .03
Reasoning Through Language Arts	.91 - .97	<.01 - .09	<.01 - .07
Science	.83 - .98	.01 - .10	<.01 - .17
Social Studies	.83 - .92	.01 - .17	<.01 - .09

Table 8. Classification Consistency in GED® English- and Spanish-Language Test Forms

Content Area	Proportion of Consistent Decisions	Chance Proportion of Correct Classification	Kappa	Proportion of Misclassification
Mathematical Reasoning	.90 - .98	.50 - .69	.80 - .96	.02 - .10
Reasoning Through Language Arts	.88 - .95	.50 - .62	.68 - .88	.05 - .12
Science	.79 - .95	.52 - .64	.45 - .90	.05 - .21
Social Studies	.78 - .88	.51 - .62	.52 - .75	.12 - .22

Chapter 8: Validity Evidence

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) provide a set of industry standards for test development and use. The opening paragraph (Chapter 1, Validity) 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. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself. When test scores are interpreted in more than one way…each intended interpretation must be validated. Statements about validity should refer to particular interpretations for specified uses. (p. 11)

An ideal validation is one that includes several types of evidence which, when combined, best reflect the value of a test for an intended purpose. The *Standards for Educational and Psychological Testing* (2014) suggest that test developers report several types of validity evidence, when appropriate. Specifically, evidence may be provided based on the following:

- Test content
- Response processes
- Internal structure
- Relations to other variables
- Consequences of testing

The sources of validity evidence included in this manual aim to follow these guidelines.

Evidence Based on Test Content

Passing the GED® test battery is a prerequisite to earning or receiving a credential. Thus, a crucial component of validity evidence is in demonstrating the ability of a test to represent adequately the content domain it purports to measure. Evidence based on test content is usually provided through rational defense of the test specifications and test development plan (Nunnally, 1978). Such evidence is demonstrated by the extent to which the items on the GED® test reflect the major content of a high school program of study.

Evidence of validity based on test content often rests on subjective analyses of test content made by subject-matter experts (Osterlind, 1989; Thorndike, 1982). Thus, to ensure adequate content representation of the GED® test, experts were used to develop the current test specifications and to evaluate each operational test form.

Content Validity Guidelines for Item Development

The GED Testing Service content validity guidelines are intended to help item writers and others achieve a common understanding of what constitutes high-quality, aligned items on the GED® test. The guidelines for each content area are living documents, updated periodically, to represent the ongoing, collective thinking of content specialists, item writers, review committee members, and others with a stake in the creation of a superior assessment.

Guidelines Components

Practices and Indicators

The guidelines name the global critical thinking skills needed to interact successfully with information in a given content area. These global thinking skills are called *practices.* Specific sub-skills of each practice are called *indicators.* It is the indicator skills that are assessed on the test.

Example:

Social Studies <u>Practice</u> (SSP) 7: Evaluating reasoning and evidence

- <u>Indicator</u> SSP7.a: Distinguish among fact, opinion, and reasoned judgment in primary and secondary sources.
- <u>Indicator</u> SSP.7.b: Distinguish between unsupported claims and informed hypotheses grounded in social studies evidence.

Indicator Explanation

The primary purpose of the explanation for each indicator is to isolate the construct: that is, to make plain the proficiency, ability, or skill that items written to the indicator should assess. Clarity about the construct to be assessed is essential to creating valid and reliable items.

Explanations are usually presented in two parts. In all cases, the explanation describes the action(s) that items aligned to the indicator require of test-takers. In most cases, the explanation also describes what items written to the indicator should **not** do. Such information helps to further isolate the construct by (a) identifying and warning against approaches to creating items that describe a task similar to, but not the same as, what is required by the indicator, (b) representing common misunderstandings of the indicator, or (c) inverting the indicator (e.g., the indicator requires test-takers to identify multiple causes for a given effect, but the item does the opposite, providing the causes and asking test-takers to name the effect).

Key Definitions

Terms used to describe the complex thinking required on the GED® test are subject to nuanced interpretation. For the sake of clarity, explanations throughout the document are supported by the definitions of key terms used by GED Testing Service.

Other Indicators

Because clearly distinguishing the skills assessed by each indicator is critically important to content validity, most indicator pages have a section entitled "Other Indicators," which emphasizes and adds to the distinctions made in the Indicator Explanation. This section serves to further delineate the differences between the indicator in question and similar indicators that are often conflated during the item development process.

Stimulus Requirements

The Stimulus Requirements sections of the document describe attributes (e.g., length, textual vs. visual, authentic vs. commissioned) that stimuli for a given indicator must or must not possess.

Examples

Depending on the indicator, examples may illustrate a variety of approaches for assessing the skill at hand. In all cases, examples help bring the abstractions in the indicator explanations to life, thus providing clarity to item writers and reviewers.

Evidence Based on Internal Structure

GED Testing Service reports a single standard score for each test. This score reporting structure assumes that each test score represents a single construct and all items on that test measure this same construct. When a single construct underlies the responses to the items on a test, we describe that test as being *unidimensional.* An important component in making a validity argument for test scores is assessing the internal structure, or dimensionality, of the test.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) also indicate that test developers must assess the quality of test items in terms of fairness. In other words, it is incumbent upon the test developer to ensure that, within reason, the likelihood of producing a correct answer does not depend on the characteristics of an examinee that are external to the measure of interest. Items that result in differential likelihoods of success for different subgroups are described as having *differential item functioning* (DIF). However, final judgment as to whether an item is biased toward one group over another is relegated to a panel of expert reviewers.

Item-Total Correlations

All items on the GED® test are expected to measure a single, dominant construct. The scores on each item should therefore correlate with the total score. Items with higher correlations to the total score provide more discrimination between higher- and lower-ability test-takers. Therefore, higher correlations are preferred. During item tryout processes, items with low correlations are usually removed from the operational item pool. The distribution of point-biserial correlations is presented by content area test in Table 9.

Table 9. Distribution of Item-Total Correlations, by Content Area Test, GED®

Content Area	Distribution of Point-Biserial Correlations			
	0 to .29	**.30 to .39**	**.40 to .49**	**.50+**
Mathematical Reasoning	.32	.32	.30	.06
Reasoning Through Language Arts	.39	.41	.17	.04
Science	.43	.43	.11	.03
Social Studies	.51	.40	.09	--

Equivalence of Forms

Each test-taker must be assessed and measured according to the same construct, regardless of which test form is administered. The construct is defined within the test purpose statement and operationalized by the content and test specifications. If we conceptualize the test purpose statement and content specifications as the blueprints, then we can think of the test items as the building blocks. In that sense, the building blocks must be consistent across forms in order to maintain across-form construct equivalence.

GED Testing Service underwent a rigorous test development process. During that process, the test specifications and definitions were extensively refined such that items could be written and mapped to fine-grained indicators, assessment targets, and content standards. These items were then field tested, and surviving items were put onto a common scale using an item response model. The forms assembly process utilized a test characteristic curve such that all forms met similar statistical properties. The test blueprints ensured that all forms met the same content specifications. Some descriptive statistics for the content area tests are provided in Table 10. Specifically, the average test score, standard deviation, and minimum and maximum scores are provided for six operational test forms. The two rightmost columns show the minimum and maximum values for the six estimates.

Table 10. Descriptive Statistics for Content Area Test Scores, GED®

Content Area	Statistic	Form						Min	Max
		1	2	3	4	5	6		
Mathematical Reasoning	Mean	151	150	151	151	151	151	150	151
	SD	10	11	10	10	10	10	10	11
	Min	100	100	100	100	100	100	100	100
	Max	200	200	200	198	197	200	197	200
Reasoning Through Language Arts	Mean	155	155	156	154	155	155	154	156
	SD	10	10	9	9	9	9	9	10
	Min	100	100	100	120	113	100	100	120
	Max	200	200	200	193	189	196	189	200
Science	Mean	153	155	154	154	156	153	153	156
	SD	9	9	9	9	9	9	9	9
	Min	100	100	100	100	100	100	100	100
	Max	188	195	189	200	194	190	188	200
Social Studies	Mean	152	154	153	155	153	153	152	155
	SD	11	10	10	10	10	10	10	11
	Min	100	100	100	100	100	100	100	100
	Max	200	200	200	195	200	196	195	200

Dimensionality

In the field of psychometrics, we often refer to the underlying construct, or constructs, measured by a test form. In the case of the GED® test, we maintain that there is a single underlying construct being measured by each content area test. This assumption is testable using confirmatory factor analyses.

Data from three English-language, operational forms for each content area test were used to assess dimensionality. Specifically, the data for each test form were split by taking a simple random sample (without replacement) of 7,500 first-time test-takers. This sample was set aside for additional dimensionality assessment, if needed. The remaining sample was used to fit a single-factor model to the item response data. M*plus* version 7.2 (Muthén & Muthén, 2014) was used to estimate the model parameters.

For this study, a chi-square test was used to assess the fit of the single-factor model to the data. Rejecting the null hypothesis would be evidence that the model did not fit the data well. However, the chi-square test is sensitive to sample size in that the Type I error increases as sample size increases (Cheung & Rensvold, 2002). Thus, investigations of model-data fit were supplemented by summary fit statistics, which summarized how well the model reproduced the population covariance matrix. M*plus* output included both the comparative fit index (CFI) and the root mean square error of approximation (RMSEA). CFI values greater than .95 and RMSEA values less than .06 (Hu & Bentler, 1999; Kline, 2005) were used to indicate good model-data fit. Robust weighted least squares (WLSMV) with delta parameterization was used to estimate the model parameters.

Table 11 provides the sample size, RMSEA, and CFI results for each content area test form. Chi-square values were statistically significant for each form and thus not reported here. However, the sample sizes likely contributed to their high values. The RMSEA are consistently less than 0.06; 90% confidence intervals provided further evidence that RMSEA estimates were acceptable. With the exception of two Mathematical Reasoning forms (1 and 3), all CFI estimates were greater than 0.95. Additional analyses (using the remaining 7,500 responses) will be performed on Mathematical Reasoning forms 1 and 3 to determine whether a secondary factor may be contributing to the responses. However, given the relatively close fit of the single-factor model, it is unlikely that a secondary factor will be meaningful.

In sum, evidence suggests that there is a single, underlying trait that is being measured by each content area test form. Additional analyses will be conducted on the Spanish-language forms when sufficient data are available.

Table 11. Confirmatory Factor Analysis Results, GED®, by Content Area Test and Form

Content Area	Form	N	Root Mean Square Error of Approximation	Comparative Fit Index
Mathematical Reasoning	1	12,274	0.030	0.933
	2	12,257	0.026	0.954
	3	12,060	0.025	0.940
Reasoning Through Language Arts	1	18,056	0.020	0.971
	2	18,388	0.020	0.971
	3	18,295	0.019	0.963
Science	1	15,678	0.019	0.972
	2	16,128	0.020	0.962
	3	15,821	0.022	0.960
Social Studies	1	16,897	0.022	0.965
	2	16,800	0.022	0.957
	3	16,799	0.022	0.952

Note: Each 90% confidence interval for RMSEA indicated the estimate was <= 0.05.

Test Bias Assessment

One of the assumptions found within GED Testing Service's validity framework is that items and tasks will be free of bias. This is a testable assumption, most often accomplished via content expert bias reviews and differential item functioning (DIF) analyses. Expert bias reviews (also referred to as Fairness Committees) are an integral piece of GED Testing Service's item development process. Each item is meticulously reviewed by a Fairness Committee prior to field testing. Items flagged for any reason are removed from the pool of field test items and discarded or revised before becoming operational. Field test data are also used to gauge whether there are any obvious issues, such as certain subgroups performing exceptionally better on an item. When such evidence is found during data review, content specialists again review the item for content issues that may contribute to DIF. Items can also be removed from operational use at this stage of review.

DIF is a statistical process whereby items are evaluated for statistical anomalies that suggest a focal group (e.g., females) is not performing similarly to a reference group (e.g., males), while controlling for ability. There are many options for assessing DIF, many of which depend upon the types of items on the test. The GED® test is a mixed-format test in the sense that there are both dichotomously and polytomously scored items on each test form.

GED Testing Service used the properties of the Rasch model to conduct a uniform DIF analysis on all content area tests. While using the Rasch model, one can analyze DIF between two groups or the difference between a specific group and the average of all groups. GED Testing Service used the former approach since the interpretation is that one group performs better than the other on a given item after controlling for ability. A free calibration was used to estimate item difficulty and person abilities. Next, the person abilities were used to anchor another calibration, while iteratively estimating an item difficulty for each group, one item at a time. All test-takers at the same raw score have the same ability level, θ. For both dichotomously and polytomously scored items, the difference in difficulty is calculated as

$$\beta_{Focal} - \beta_{Reference} = \ln\left(\frac{P_{R1}}{P_{R0}}\right) - \ln\left(\frac{P_{F1}}{P_{F0}}\right)$$

where P_{R1} and P_{F1} are the proportions of test-takers who received a correct score on the item in the reference and focal groups, respectively, and ($P_{R0} = 1\text{-}P_{R1}$) and ($P_{F0} = 1\text{-}P_{F1}$).[20] The difference between the estimates can be tested for statistical significance by dividing by the standard error of the difference.

$$t = \frac{(\beta_F - \beta_R)}{SE(\beta_F - \beta_R)}$$

DIF analyses were performed on gender (male, female), race (White/Caucasian, Black or African American), and ethnicity (Hispanic, Non-Hispanic). These demographic variables are collected at the time of registration, though respondents can opt -out of each question. Therefore, only those candidates who provided information for each variable were eligible for analysis. DIF analyses are sensitive to both sample sizes overall as well as comparative group sample sizes. For the race and ethnicity analyses, random sampling was used to select a sample of similar size to the focal groups (i.e., Black or African American and Hispanic). DIF analyses were performed at the trait level for extended response (ER) items. Only data from first-test experience were included.

Items without DIF are often flagged due to large sample sizes. Thus, some level of effect size is generally used to flag meaningful DIF. Items with an absolute difference in item difficulty between .43 and .63 were labeled as Level B DIF, and items with differences greater than or equal to .64 were labeled Level C. These values are synonymous with the Educational Testing Service delta scale, where 1 logit is equal to 2.35 delta points. Items that were labeled as Level B or C and statistically significant were flagged for further bias review. False positive rates were controlled using the Benjamini-Hochberg procedure (see Thissen, Steinberg, & Kuang, 2002).

[20] The terms Focal and Reference groups are traditionally used in DIF analyses. The Rasch approach used here does not necessarily rely on such designations.

Summary DIF analyses results are presented in Table 12. Specifically, the percentage of the total content-specific item pool flagged for each analysis variable and the number of Level C DIF items are listed in the table. For example, approximately 1 percent of the Mathematical Reasoning test items in the operational item pool were flagged for DIF, and none of them were at the C level. The gender and race variables accounted for more DIF items than ethnicity. For Social Studies and Reasoning Through Language Arts, the ER items accounted for a large number of the DIF items. Specifically, females performed better on the ER items than males. The reader should note that the results are nonspecific to either the focal group or the reference group. That is, the summary in Table 12 includes items flagged for DIF in either direction, such as those favoring either Hispanic or non-Hispanic, males or females, White/Caucasian or Black/African American.

Table 12. Summary of Differential Item Functioning Analyses, GED®

Content Area	Ethnicity		Gender		Race	
	Percent	Level C	Percent	Level C	Percent	Level C
Mathematical Reasoning	.01	0	.05	1	.25	10
Reasoning Through Language Arts	.01	0	.06	1	.06	1
Science	.04	0	.11	2	.22	4
Social Studies	.02	1	.12	9	.16	4

All items undergo a bias review process prior to becoming operational. DIF analysis results are statistical in nature and do not necessarily reflect bias in the items. Furthermore, DIF analyses are inherently more complex than described here, and there are various processes for analyzing DIF. As described by Andrich and Hagquist (2012), large amounts of DIF found in a single item can result in artificial DIF in the remaining items (which varies by test length) without careful control procedures. Thus, a relatively conservative approach is to re-examine all items flagged for DIF via another round of bias review, using the statistical results to guide the discussion.

At the time of this publication, the items flagged for DIF were still undergoing a second round of bias review (including both internal and external reviewers). Items deemed biased based on the results of this review process will be removed from the operational item pool.

Speededness Assessment

Exam times for each content area (averaged across all test forms) are presented in Table 13. The allotted time is 150 minutes for Reasoning Through Language Arts, 115 minutes for Mathematical Reasoning, and 90 minutes for Science and Social Studies. Note for this analysis that the minimum exam time was limited to 5 minutes to eliminate exams that were abandoned at the start. Test-takers who received extra time accommodations may be included in the data.

Table 13. GED® Exam Times, by Content Area Test

Content Area	Mean	SD	50th	75th	85th	90th	95th	Max
Mathematical Reasoning	1:26	0:23	1:30	1:48	1:51	1:53	1:53	1:55
Reasoning Through Language Arts	1:47	0:24	1:50	2:08	2:15	2:17	2:18	2:29
Science	1:13	0:15	1:18	1:27	1:28	1:28	1:28	1:30
Social Studies	1:10	0:16	1:14	1:23	1:27	1:28	1:28	1:30

Correlations Among Content Area Tests

Correlations among the content area tests are presented in Table 14. The lower half of the correlation matrix shows the raw correlation, while the upper half shows the correlation corrected for measurement error.

Table 14. Correlations among GED® Content Area Tests

	Reasoning Through Language Arts	Mathematical Reasoning	Science	Social Studies
Reasoning Through Language Arts		.66	.71	.90
Mathematical Reasoning	.57		.82	.72
Science	.71	.69		.87
Social Studies	.74	.60	.70	

Chapter 9: Accommodations for GED® Test-Takers with Disabilities

The purpose of accommodations is to provide candidates with full access to the GED® test. However, accommodations are not a guarantee of improved performance or test completion. GED Testing Service provides reasonable and appropriate accommodations to individuals with documented disabilities who demonstrate a need for accommodations. Several accommodations are available for the 2014 GED® test. These accommodations include the following, but are not limited to:

- Extended testing time (e.g., 25%, 50%, 100% additional time)
- Extra breaks
- A private testing room
- A paper version of the test for test takers that have a medical condition that precludes them from using a computer screen.
- An audio version of the test (provided through the computer using Job Access With Speech [JAWS] screen reading technology integrated into the computer test driver)
- Presentation of the test material on computer in an appropriate point size as determined by an individual test-taker using a "zoom" feature known as Zoom Text
- Braille
- Talking calculator for visually impaired test-takers
- Scribe and/or reader for test-takers with a variety of physical conditions that prohibit test-takers from reading or responding on their own

It should also be noted that any test taker can select from a range of background and text display combinations and enlarge the font up to 20 point which provides the most comfortable and appropriate computer testing environment. Test accommodations are individualized and considered on a case-by-case basis. Consequently, no single type of accommodation (e.g., extra time or a paper test) would necessarily be appropriate for all individuals with disabilities. Simply demonstrating that an individual meets diagnostic criteria for a particular disorder does not mean that the person is automatically entitled to accommodations.

GED Testing Service has a formal process in place for applying for an accommodation. A completed application and supporting documentation must be submitted for review. These request form and support documentation guidelines are on the GED Testing Service website here: http://www.gedtestingservice.com/testers/computer-accommodations.

1. Accommodations request forms:
    a. Intellectual Disabilities (ID)
    b. Learning and Other Cognitive Disabilities (LCD)
    c. Attention-Deficit/Hyperactivity Disorder (ADHD)
    d. Psychological and Psychiatric Disorders (EPP)
    e. Physical Disorders and Chronic Health Conditions (PCH)
    f. Request for Testing Accommodations Appeal
    g. Request for Extension
2. Supporting Documentation and Guidelines:
    a. Evaluators: Intellectual Disabilities
    b. Evaluators: Learning and Other Cognitive Disorders
    c. Evaluators: Attention-Deficit/Hyperactivity Disorder
    d. Evaluators: Psychological and Psychiatric Disorders
    e. Evaluators: Long-Term Physical Disabilities and Chronic Health Conditions

Accommodations Procedures Summary

When test-takers register online for the GED® test on computer, the following process applies.

1. The test-taker registers online for the GED® test on computer.
2. During the registration process, the test-taker is asked about the need for accommodations or modified testing conditions.
3. If the test-taker has a need for accommodations, then the following steps occur:
    a. The test-taker is directed to the GED Testing Service accommodations website (www.gedtestingservice.com/accommodations) for instructions regarding applying for accommodations.
    b. The test-taker is told that scheduling of an accommodated seat cannot take place until the accommodations approval process has been completed.
    c. The test-taker is informed to expect up to 30 days for review of the accommodations request.

The test-taker then follows the steps for applying for accommodations, as outlined on the website specified above, including completing the Accommodations Request Form and gathering supporting documentation. The test-taker submits the Accommodations Request Form and supporting documentation according to the directions on the website. GED Testing Service reviews the test-taker's submission and makes the accommodations determination. If GED Testing Service needs additional information pertaining to the test-taker's accommodation request, GED Testing Service contacts the test-taker for this purpose. The accommodations decision is entered into the GED Testing Service registration system and the test-taker is directed to a dedicated accommodation scheduling phone number to schedule their accommodated tests.

Accommodations for GED® Test-takers Who Are Blind or Have Visual Impairments

There are a number of testing alternatives for test-takers who are blind or who have visual impairments that could potentially hamper access to the GED® test. These include, but are not limited to, the following:

1. Use of a scribe for recording responses
2. Use of a human reader
3. Use of Screen magnification of the GED® test on computer
4. Use of a screen overlay for the GED® test on computer
5. Use of a talking calculator
6. Use of the Braille GED® test

In unusual situations, more individualized accommodations can be arranged, but this may require more than the standard 30 days to review and prepare for such accommodations.

Accommodations for Test-takers Who Are Deaf or Hard of Hearing

Because the GED® tests are written (not oral), there are minimal requirements for most test-takers who are deaf or hard of hearing. In no situation may the GED® test be translated or interpreted into any other language, such as American Sign Language, which would fundamentally alter the nature of the test. If a test-taker is deaf or hard of hearing, the following adaptation may be approved:

1. Instead of listening to the test administrator or proctor say the instructions aloud (prior to the start of the test), the test administrator may provide the instructions to the test-taker in written form.

Accommodations for Test-takers Who Have ADHD

Test-takers who have Attention-Deficit/Hyperactivity Disorder (with or without hyperactivity/impulsivity; hereafter called ADHD) may request testing accommodations. In order to be approved, a test-taker with ADHD (like any other disorder) must not only demonstrate that diagnostic criteria for the disorder have been met, through appropriate documentation, but also that the disorder rises to the level of a *disability* as defined by applicable federal law. That is, test-takers must provide evidence that they are substantially limited in a major life activity, not merely that they have symptoms of inattention or distractibility. After a test-taker provides evidence that he or she is *disabled* as defined by law, the test-taker must show that the requested accommodations are reasonable and necessary in order to provide access to the GED® test. Test-takers with ADHD may benefit most from extra “refocusing” breaks, which are supervised, and/or testing in a private or distraction-reduced room. In most cases, for test-takers who have trouble sustaining their attention over time, and/or for whom attention begins to wane over time, the use of extended testing time is contraindicated. That is, test-takers who have trouble sustaining their attention over time, and their evaluators, should carefully consider the logic of any request to dramatically lengthen the seat time.

Chapter 10: Supplemental Materials for Test Preparation

Since 2012, GED Testing Service has been producing an extensive set of materials and resources in support of the new GED® test. These materials include manuals, paper-based and computer-based learning tools, practice tests in a variety of formats, recorded webinars, self-study materials, online learning courses, and more. Below is a brief summary of three key products that have been prepared to help test-takers and educators prepare for the GED® test. These products are the GED® Test Tutorial, the GED® Test Item Samplers, and GED Ready®: The Official Practice Test.

GED® Test Tutorial

GED® Testing Service developed a computer skills tutorial for the 2014 GED® Test. As part of that tutorial development, GED® Testing Service commissioned a usability study of the tutorial. The purpose of the GED® Test Tutorial was to provide potential test-takers with information about the necessary computer skills for the computer-based version of the GED® test. This study was intended to research and establish that the GED® Test Tutorial is accessible, usable, and useful for the GED® test-taking population.

The GED® testing population has wide variation in both background and demographic characteristics. Many GED® test-takers have been out of school for many years, while others have only recently left high school. In order to make the tutorial accessible to such a diverse population, GED® Testing Service commissioned a usability study.

The usability study for the GED® Test Tutorial was held across three rounds, with substantial revisions and improvements made between each round. Each usability session was conducted individually, and the *think aloud* usability method was employed. All sessions were held at the Hubbs Center for Lifelong Learning in Saint Paul, Minnesota. A breadth of usability participants was sought, with the participants' readiness to take the GED® test as the most important selection criterion.

Several aspects of the GED® Test Tutorial were successful from the beginning. The participants were almost universally able to use all of the question types presented in Section 2, Question Types (i.e., multiple choice, fill-in-the-blank, drop-down, hot spot, and drag-and-drop). Furthermore, participants were largely successful in computer-use tasks addressed throughout the tutorial, such as accessing page tabs, finding buttons onscreen, opening exhibits, and interpreting pop-up messages from the delivery software.

There were three primary areas of difficulty with the initial tutorial. The difficulties were in the areas of Navigator, Editing Tools, and Mathematics Resources. Most of the substantive revisions were aimed at resolving these usability challenges, and these revisions yielded substantially improved usability.

By the end of the study, almost all participants understood the purpose of Navigator, and they were far better positioned to use it successfully. In addition, even participants with little or no prior word processing experience were able to follow the instruction regarding Editing Tools. Finally, participants were able to use and understand the various Mathematics Resources to a far greater extent than had been possible initially.

It is important to note that taking a computer-based test requires a broad set of software use skills, as well as other background skills such as literacy and facility with the English language. The participants in this study were selected primarily based on their "preparedness" for taking the GED® test, as this level of readiness was seen as encapsulating many of the background skills needed to benefit from the tutorial. It is anticipated that future users of the GED® Test Tutorial who are prepared to take the GED® test will also have these necessary background skills and will be able to use the tutorial successfully.

However, future tutorial users who are unable to use a computer keyboard will need additional instruction in typing before they are able to benefit from the Editing Tools section of the tutorial (Section 3), or to successfully take any GED® test that includes short answer or extended response questions. In a similar fashion, tutorial users with little training in mathematics are likely to find the Mathematics Resources section of the tutorial (Section 4) to be overly challenging. Finally, those tutorial users who have *low* or *very low* computer experience are still likely to need additional support beyond the tutorial itself, perhaps in the form of participation in Adult Basic Education programs.

GED® Test Item Samplers

In order to help educators and other stakeholders better understand how the 2014 GED® test Assessment Targets are reflected in test items, GED Testing Service has prepared Item Samplers for each content area test. The Item Samplers are available on the GED Testing Service website in three formats:

1. An interactive version for viewing on the web
2. An interactive version for download and viewing on a test-taker's computer at any time
3. A PDF version designed for users who want or need a paper-based view of the Item Samplers

The Item Samplers provide users with examples of test items that represent the full range of item types that appear on each content-area test in the 2014 GED® test. For example, the Reasoning Through Language Arts (RLA) Item Sampler includes items that demonstrate how the various item types appear in the context of that content-area test. Because short answer items will no longer appear on the RLA exam in order to minimize testing time, the Item Sampler for RLA no longer includes an item type in that category.

Second, the Item Samplers show how some of the 2014 GED® test Assessment Targets can be translated into test items. By examining the Answer Explanation tab for each item, a user can see which Assessment Target is being measured by the item, and then trace the target back to the instructional career- and college-readiness standard that is being measured.

Finally, the interactive versions of the Item Samplers provide the look, feel, and core functionality of the testing interface that test-takers can see when they take the 2014 GED® test on computer. The Item Samplers have been built with different software than the 2014 GED® test, however, for ease of distribution and accessibility by a broad range of stakeholders. Even though the Samplers have the essential functionality, look, and feel of the test, the actual test contains additional functionality that is not present in the Item Samplers, such as support for accommodations, a functioning onscreen calculator, and so on. Of course, the Item Samplers also contain functionality that does not appear on the actual 2014 GED® test, such as the Answer Rationale button.

Users of the Item Samplers are cautioned not to make inferences about the 2014 GED® test based solely on the Item Samplers.

First, the Item Samplers are not intended to demonstrate the full range of difficulty of the 2014 GED® test. The original July 2012 Item Samplers were published prior to the 2012 GED Testing Service Field Test, in which items that would appear on the actual 2014 GED® test were systematically administered to candidates to determine their difficulty and ability to discriminate performance among test-takers of varying skill levels. Because of this, the items in the samplers should not be interpreted by users as representative of the full range of item difficulty on the 2014 GED® test.

Second, because the Item Samplers are intended to provide examples of all the different items types that can appear on each content area test, the samplers do not reflect the actual distribution of items across the various item types. Multiple choice items still represent the majority of items on each content area test. As multiple choice is such a familiar item type to educators and test-takers, fewer of those items have been included in the Item Samplers.

Finally, as a result of field testing conducted in the summer of 2012, the revised November 2012 publication of the Item Samplers now includes 16 additional multiple choice items (four items in each content area) that have been field tested and that:

1. Reflect the 2014 GED® test Assessment Targets
2. Show additional examples of multiple choice items
3. Provide sample items that are easier to more moderate in difficulty, thereby depicting items that are more representative of the range represented on the 2014 test

In short, the Item Samplers are a convenient way to demonstrate all of the item types and functionality, but the samplers are not fully representative of content or difficulty of the 2014 GED® test. The Item Samplers' purpose is simply to show which item types are employed in each content area.

GED Ready®: The Official Practice Test

GED Ready®: The Official Practice Test (GED Ready®) is a computer-based practice test that is approximately 50 percent of the length of the operational 2014 GED® test. Like the operational test, GED Ready® is standardized and placed on the same scale as the operational test. These attributes allow GED Ready® to provide information to test-takers on their likelihood of passing the 2014 GED® test, descriptive information about their skills as demonstrated on the test, and what they will need to master to move to the next level of performance on the test. Early results show that 95 percent of those who score in the “Likely to Pass” zone of GED Ready® are going on to pass the operational GED® test.

GED Ready® is administered over the Internet via a secure connection and provides immediate, individualized score reports to test-takers, along with information that will allow adult educators to score the constructed response items and provide feedback to test-takers. The test launched with two forms initially, but additional forms continue to be added as appropriate. In correctional settings, GED Ready® is available in an alternative version not requiring Internet access, delivered in the same way the operational GED® test is delivered. GED Ready® is available for purchase by test-takers through the *MyGED*™ portal. Institutions are also able to purchase GED Ready® for distribution to their students through arrangements with the GED® publisher network. GED Testing Service works with over 15 publishers to resell GED Ready® to their customers. By allowing these publishers to distribute GED Ready®, GED Testing Service provides greater flexibility for adult education programs to purchase GED Ready® through existing publishers from which they already buy other test preparation materials. Publishers offering GED Ready® are listed at www.gedtestingservice.com/educators/2014publishers.

Purposes of GED Ready®

There are two primary purposes for GED Ready®. The first purpose is to provide feedback about future performance on the GED® test. GED Testing Service uses an item response theory model to predict test-taker performance on the full-length, operational GED® test based on the test-taker's performance on GED Ready®, *excluding the short answer and extended response items*.

The second purpose of GED Ready® is to provide a practice opportunity for test-takers. This practice opportunity allows adults to experience the types of items that will appear on the operational GED® test as well as the how the items are displayed, responded to, and navigated to. Thus, GED Ready® administers all item types, including multiple choice, fill-in the-blank, drop-down, hot spot, drag-and-drop, short answer, and extended response items.

In connection with the first purpose of GED Ready®, the exam provides test-takers with an indication of how they might perform on the full-length, operational GED® test. In addition to a test score, test-takers are classified into one of three performance categories or zones. The *Standards for Psychological and Educational Testing* state that when classification decisions are made based on test scores, “…estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.” (p. 46). (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

GED Testing Service made the strategic policy decision to offer computer-based testing as the primary mode of administration. This decision was based on several factors, including (a) taking advantage of new, innovative item types, (b) leveraging computer-based scoring techniques, (c) using rapid score reporting, and (d) following a specific recommendation from the Technical Advisory Committee to avoid potential cross-mode comparability. Thus, the GED® test is a single-mode test, with very few exceptions being offered for accommodations purposes. Accordingly, at this time, GED Testing Service offers only a computer-based version of GED Ready®. However, given the market need, we are now in the process of creating forms for GED Ready® for paper, Braille, and screen reader in both English and Spanish.

GED Ready® Test Characteristics

GED Ready® is essentially an abbreviated version of the operational GED® test. It is a battery comprising four content area tests: Reasoning Through Language Arts (RLA), Mathematical Reasoning, Science, and Social Studies. Each test is a practice version of the full-length, operational GED® test of the same content area.

All GED Ready® test items are developed simultaneously with those for the operational GED® test and are drawn from the same item pool. Furthermore, GED Ready® contains all of the same item types and formats of the operational test. The main difference between GED Ready® and the GED® test—aside from test length—involves scoring constructed response items. Both RLA and Social Studies have extended response items, and Science has short answer items. Although these items do appear on GED Ready®, they are not subject to standardized, professional human scoring.[21]

GED Ready® and GED® test scores are reported on the same scale, which ranges from 100 to 200. Because the purpose of GED Ready® is to provide feedback on future performance on the GED® test, GED Testing Service created the following performance reporting zones to aid score interpretation: (1) Not Likely to Pass (the "red" zone), (2) Too Close to Call (the "yellow" zone), and (3) Likely to Pass (the "green" zone).

Reliability of GED Ready®

Reliability refers to the consistency, or stability, of the test scores over replications and is inversely related to the amount of measurement error associated with test scores. There are multiple ways to measure reliability. One of the strongest methods of assessing test score reliability is to administer parallel forms to the same group of test-takers on multiple occasions (assuming those occasions represent real-world scenarios). This method is often referred to as *alternate forms reliability.* When test scores are used to classify test-takers into discrete categories (e.g., categories that predict the likelihood of future success), then the reliability of those classifications is referred to as *decision consistency.*

[21] These items will not yield an immediate score on GED Ready® and thus cannot be used in the prediction. Users will have access to a complete set of example test-taker responses representing different scores on those items and explanations of scores.

Alternate Forms Reliability

Table 15 below presents the scaled score correlations between two forms of GED Ready® that were administered during the 2013 Standardization and Norming study. The sample sizes include only those participants who were administered both forms as part of the spiral. The correlations range from .73 to .88 across the content area tests. The correlation of true scores can be estimated by removing the measurement error. The rightmost column of Table 15 shows the true score correlations, which range from .94 to 1.0.[22]

Table 15. Alternate Forms Reliability Estimates, GED Ready®

Content Area	Sample Size	Correlation	Corrected for Attenuation
Reasoning Through Language Arts	181	.73	0.94
Mathematical Reasoning	186	.88	1.00
Science	182	.86	1.00
Social Studies	186	.73	0.95

Correlations between scaled scores on GED Ready® and a single GED® test form (Form A) were also calculated. In general, these correlations, which are presented in Table 16, were similar across the two GED Ready® forms. In all but one case, the correlation between two GED Ready® forms (Table 16) was higher than those between GED Ready® and the GED® test. Once corrected for attenuation, the correlations between GED Ready® and the GED® test true scores were lower compared to those between GED Ready® forms. For both RLA and Social Studies, the disattenuated correlations were not as high as those for Mathematical Reasoning and Science. This is likely because RLA and Social Studies both include the extended response item on the GED® test.

Table 16. Correlations Between GED Ready® and the GED® Test

Content Area	Correlation with Form A		Corrected for Attenuation	
	Form 1	**Form 2**	**Form 1**	**Form 2**
Reasoning Through Language Arts	.72	.71	.83	.87
Mathematical Reasoning	.82	.86	.92	.95
Science	.76	.75	.95	.91
Social Studies	.69	.74	.83	.89

[22] Note that correlations for raw scores were slightly higher and those corrected for attenuation were all estimated at 1.0. The difference likely has to do with some raw scores resulting in the same scaled score, thus reducing some of the score variability.

Decision Consistency

GED Ready® test scores are used to classify test-takers into one of three performance categories. An important estimate of reliability, then, is the consistency of test-taker classifications across administrations of the same test. As mentioned, two forms of GED Ready® were administered during the 2013 Standardization and Norming Study. Therefore, one can measure the consistency of observed score classifications across these two forms. *Kappa* is a statistic that measures the agreement between two forms. *Quadratic (weighted) kappa* is an extension of the kappa statistic that penalizes more for non-adjacent classifications. For both statistics, a value of zero would indicate agreement is due solely to chance, whereas a value of 1.0 represents perfect agreement.

Results of the decision consistency analyses (including all three performance categories simultaneously) are presented in Table 17. Kappa statistics range from .50 to .70. Weighted kappa statistics range from .59 to .78.

Table 17. Decision Consistency Estimates for GED Ready®

Content Area	*N*	Kappa	Weighted Kappa
Reasoning Through Language Arts	181	.55	.72
Mathematical Reasoning	186	.70	.78
Science	182	.64	.74
Social Studies	186	.50	.59

Conditional Standard Errors of Measurement

Conditional standard errors of measurement (CSEM) represent the standard deviations of measurement errors at a given score level. Table 18 presents the CSEM for scaled scores within the vicinity of both GED Ready® performance zone cut scores. The GED® test score scale ranges from 100 to 200, with a scaled score of 150 representing the Passing Standard and a scaled score of 170 representing the GED® with Honors Performance Level.

Table 18. Conditional Standard Errors of Measurement for GED Ready®

Content Area	Form 1		Form 2	
	Passing Standard	**GED Score with Honors**	**Passing Standard**	**GED Score with Honors**
Reasoning Through Language Arts	5	5	5	5
Mathematical Reasoning	5	5	5	5
Science	6	6	6	5
Social Studies	7	6	7	6

Predictive Validity

Table 19 below provides the percentage of adult testers who have taken a GED Ready® exam prior to their first operational test. Many additional adults have taken GED Ready® but either have not taken the operational test or took GED Ready® after their first operational test experience.

Table 19. Percentage of Adult Testers Who Have Taken GED Ready®

Content Area	Percentage
Reasoning Through Language Arts	35%
Mathematical Reasoning	32%
Science	38%
Social Studies	38%

Table 20 shows the percentage of adult testers who passed the operational GED® test given their performance on GED Ready®. The pass rate is given by content area and performance zone. As can be seen in the table, the percentage of passers is extremely high for those who scored in the green zone on GED Ready®.

Table 20. Percentage of Adult Testers Who Passed the GED® Test, Given GED Ready® Performance

Content Area	GED Ready® Zone	Passed
Reasoning Through Language Arts	Red	28%
	Yellow	73%
	Green	97%
Mathematical Reasoning	Red	20%
	Yellow	56%
	Green	92%
Science	Red	41%
	Yellow	69%
	Green	93%
Social Studies	Red	30%
	Yellow	59%
	Green	90%

References

American Council on Education. (1999). *Alignment of national and state standards: A report by the GED Testing Service.* Washington, D.C.: Author.

American Council on Education. (1964). *Examiners manual for the Tests of General Educational Development, high school level.* Washington, D.C.: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics, 37*(3), 387–416.

Brennan, R. L. (2004). *Manual for BB-Class: A computer program that uses the beta-binomial model for classification consistency and accuracy (CASMA Rep. No. 9)* [Computer software manual]. Retrieved from www.education.uiowa.edu/centers/casma/computer-programs

Cheung, G., & Rensvold, R. (2002). Evaluating goodness of fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–245.

Feldt, L. S. (2002). Estimating the internal consistency reliability of tests composed of testlets varying in length. *Applied Measurement in Education, 15*(1), 33–48.

GED Testing Service. (2011). *2011 Annual statistical report on the GED® Test.* Retrieved from http://www.gedtestingservice.com/uploads/files/4176ab251366d3ccfb4e94a9a888e67a.pdf

GED Testing Service. (2013). *2012 Stand-alone field test: Technical manual.* Washington, DC: Author.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 443–470). Westport, CT: Praeger Publishers.

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

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger Publishers.

Kline, R. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analysis. *Educational Measurement: Issues and Practice, 19*(4), 9–15.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998). *The bookmark standard setting procedure: Methodology and recent implementations.* Paper presented at the National Council for Measurement in Education annual meeting, San Diego, CA.

Lindquist, E. F. (1944, October). The use of tests in the accreditation of military experience and in the educational placement of war veterans. *Educational Record,* 357–376.

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

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.

Mullane, L. (2001). *Bridges of opportunity: A history of the Center for Adult Learning and Educational Credentials.* Washington, D.C.: American Council on Education.

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Osterlind, S. J. (1989). *Constructing test items.* Norwell, MA: Academic Press.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg Procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77–83.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Whitney, D. R., & Patience, W. M. (1981). *Work rates on the GED Tests: Relationships with examinee age and test time limits.* (GEDTS Research Studies, No. 6). Washington, D.C.: American Council on Education.

Williamson, D.M., Xi, X., & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*, 2–13.

Appendix A: Reasoning Through Language Arts Scoring Rubric

Score	Description
Trait 1: Creation of Arguments and Use of Evidence	
2	- generates text-based argument(s) and establishes a purpose that is connected to the prompt - cites relevant and specific evidence from source text(s) to support argument (may include few irrelevant pieces of evidence or unsupported claims) - analyzes the issue and/or evaluates the validity of the argumentation within the source texts (e.g., distinguishes between supported and unsupported claims, makes reasonable inferences about underlying premises or assumptions, identifies fallacious reasoning, evaluates the credibility of sources, etc.)
1	- generates an argument and demonstrates some connection to the prompt - cites some evidence from source text(s) to support argument (may include a mix of relevant and irrelevant citations or a mix of textual and non-textual references) - partially analyzes the issue and/or evaluates the validity of the argumentation within the source texts; may be simplistic, limited, or inaccurate
0	- may attempt to create an argument OR lacks purpose or connection to the prompt OR does neither - cites minimal or no evidence from source text(s) (sections of text may be copied from source) - minimally analyzes the issue and/or evaluates the validity of the argumentation within the source texts; may completely lack analysis or demonstrate minimal or no understanding of the given argument(s)

Non-scorable Responses (Score of 0/Condition Codes)

- Response exclusively contains text copied from source text(s) or prompt
- Response shows no evidence that test-taker has read the prompt or is off-topic
- Response is incomprehensible
- Response is not in English
- Response has not been attempted (blank)

Score	Description
Trait 2: Development of Ideas and Organizational Structure	
2	- contains ideas that are well developed and generally logical; most ideas are elaborated upon - contains a sensible progression of ideas with clear connections between details and main points - establishes an organizational structure that conveys the message and purpose of the response; applies transitional devices appropriately - establishes and maintains a formal style and appropriate tone that demonstrate awareness of the audience and purpose of the task - chooses specific words to express ideas clearly
1	- contains ideas that are inconsistently developed and/or may reflect simplistic or vague reasoning; some ideas are elaborated upon - demonstrates some evidence of a progression of ideas, but details may be disjointed or lacking connection to main ideas - establishes an organization structure that may inconsistently group ideas or is partially effective at conveying the message of the task; uses transitional devices inconsistently - may inconsistently maintain a formal style and appropriate tone to demonstrate an awareness of the audience and purpose of the task - may occasionally misuse words and/or choose words that express ideas in vague terms
0	- contains ideas that are insufficiently or illogically developed, with minimal or no elaboration on main ideas - contains an unclear or no progression of ideas; details may be absent or irrelevant to the main ideas - establishes an ineffective or no discernable organizational structure; does not apply transitional devices, or does so inappropriately - uses an informal style and/or inappropriate tone that demonstrates limited or no awareness of audience and purpose - may frequently misuse words, overuse slang or express ideas in a vague or repetitive manner

Non-scorable Responses (Score of 0/Condition Codes)

- Response exclusively contains text copied from source text(s) or prompt
- Response shows no evidence that test-taker has read the prompt or is off-topic
- Response is incomprehensible
- Response is not in English
- Response has not been attempted (blank)

Score	Description
Trait 3: Clarity and Command of Standard English Conventions	
2	- demonstrates largely correct sentence structure and a general fluency that enhances clarity with specific regard to the following skills: 1) varied sentence structure within a paragraph or paragraphs 2) correct subordination, coordination and parallelism 3) avoidance of wordiness and awkward sentence structures 4) usage of transitional words, conjunctive adverbs and other words that support logic and clarity 5) avoidance of run-on sentences, fused sentences, or sentence fragments - demonstrates competent application of conventions with specific regard to the following skills: 1) frequently confused words and homonyms, including contractions 2) subject-verb agreement 3) pronoun usage, including pronoun antecedent agreement, unclear pronoun references, and pronoun case 4) placement of modifiers and correct word order 5) capitalization (e.g., proper nouns, titles, and beginnings of sentences) 6) use of apostrophes with possessive nouns 7) use of punctuation (e.g., commas in a series or in appositives and other non-essential elements, end marks, and appropriate punctuation for clause separation) - may contain some errors in mechanics and conventions, but they do not interfere with comprehension; overall, standard usage is at a level appropriate for on-demand draft writing.
1	- demonstrates inconsistent sentence structure; may contain some repetitive, choppy, rambling, or awkward sentences that may detract from clarity; demonstrates inconsistent control over skills 1-5 as listed in the first bullet under Trait 3, Score Point 2 above - demonstrates inconsistent control of basic conventions with specific regard to skills 1 – 7 as listed in the second bullet under Trait 3, Score Point 2 above - may contain frequent errors in mechanics and conventions that occasionally interfere with comprehension; standard usage is at a minimally acceptable level of appropriateness for on-demand draft writing.
0	- demonstrates consistently flawed sentence structure such that meaning may be obscured; demonstrates minimal control over skills 1-5 as listed in the first bullet under Trait 3, Score Point 2 above - demonstrates minimal control of basic conventions with specific regard to skills 1 – 7 as listed in the second bullet under Trait 3, Score Point 2 above - contains severe and frequent errors in mechanics and conventions that interfere with comprehension; overall, standard usage is at an unacceptable level for on-demand draft writing. OR - response is insufficient to demonstrate level of mastery over conventions and usage

* Because test-takers will be given only 45 minutes to complete Extended Response tasks, there is no expectation that a response should be completely free of conventions or usage errors to receive a score of 2.

Non-scorable Responses (Score of 0/Condition Codes)

- Response exclusively contains text copied from source text(s) or prompt
- Response shows no evidence that test-taker has read the prompt or is off-topic
- Response is incomprehensible
- Response is not in English
- Response has not been attempted (blank)

Appendix B: Social Studies Scoring Rubric

Score	Description
Trait 1: Creation of Arguments and Use of Evidence	
2	- generates a text-based argument that demonstrates a clear understanding of the relationships among ideas, events, and figures as presented in the source text(s) **and** the historical contexts from which they are drawn - cites relevant and specific evidence from primary and secondary source text(s) that adequately supports an argument - is well-connected to both the prompt and the source text(s)
1	- generates an argument that demonstrates an understanding of the relationships among ideas, events, and figures as presented in the source text(s) - cites some evidence from primary and secondary source texts in support of an argument (may include a mix of relevant and irrelevant textual references) - is connected to both the prompt and the source text(s)
0	- may attempt to create an argument but demonstrates minimal or no understanding of the ideas, events and figures presented in the source texts or the contexts from which these texts are drawn - cites minimal or no evidence from the primary and secondary source texts; may or may not demonstrate an attempt to create an argument. - lacks connection either to the prompt or the source text(s)

Non-scorable Responses (Score of 0/Condition Codes)

- Response exclusively contains text copied from source text(s) or prompt
- Response demonstrates that the that test-taker has read neither the prompt nor the source text(s)
- Response is incomprehensible
- Response is not in English
- Response has not been attempted (blank)

Trait 2: Development of Ideas and Organizational Structure	
1	- Contains a sensible progression of ideas with understandable connections between details and main ideas - Contains ideas that are developed and generally logical; multiple ideas are elaborated upon - Demonstrates appropriate awareness of the task
0	- Contains an unclear or no apparent progression of ideas - Contains ideas that are insufficiently developed or illogical; just one idea is elaborated upon - Demonstrates no awareness of the task

Non-scorable Responses (Score of 0/Condition Codes)

- Response exclusively contains text copied from source text(s) or prompt
- Response demonstrates that the that test-taker has read neither the prompt nor the source text(s)
- Response is incomprehensible
- Response is not in English
- Response has not been attempted (blank)

Trait 3: Clarity and Command of Standard English Conventions	
1	- demonstrates adequate applications of conventions with specific regard to the following skills: 1) frequently confused words and homonyms, including contractions 2) subject-verb agreement 3) pronoun usage, including pronoun antecedent agreement, unclear pronoun references, and pronoun case 4) placement of modifiers and correct word order 5) capitalization (e.g., proper nouns, titles, and beginnings of sentences) 6) use of apostrophes with possessive nouns 7) use of punctuation (e.g., commas in a series or in appositives and other non-essential elements, end marks, and appropriate punctuation for clause separation) - demonstrates largely correct sentence structure with variance from sentence to sentence; is generally fluent and clear with specific regard to the following skills: 1) correct subordination, coordination and parallelism 2) avoidance of wordiness and awkward sentence structures 3) usage of transitional words, conjunctive adverbs and other words that support logic and clarity 4) avoidance of run-on sentences, fused sentences, or sentence fragments 5) standard usage at a level of formality appropriate for on-demand, draft writing. - may contain some errors in mechanics and conventions, but they do not interfere with understanding*
0	- demonstrates minimal control of basic conventions with specific regard to skills 1 – 7 as listed in the first bullet under Trait 3, Score Point 1 above - demonstrates consistently flawed sentence structure; minimal or no variance such that meaning may be obscured; demonstrates minimal control over skills 1-5 as listed in the second bullet under Trait 3, Score Point 1 above - contains severe and frequent errors in mechanics and conventions that interfere with comprehension OR - response is insufficient to demonstrate level of mastery over conventions and usage

*Because test-takers will be given only 25 minutes to complete Extended Response tasks, there is no expectation that a response should be completely free of conventions or usage errors to receive a score of 1.

Non-scorable Responses (Score of 0/Condition Codes)

- Response exclusively contains text copied from source text(s) or prompt
- Response demonstrates that the that test-taker has read neither the prompt nor the source text(s)
- Response is incomprehensible
- Response is not in English
- Response has not been attempted (blank)