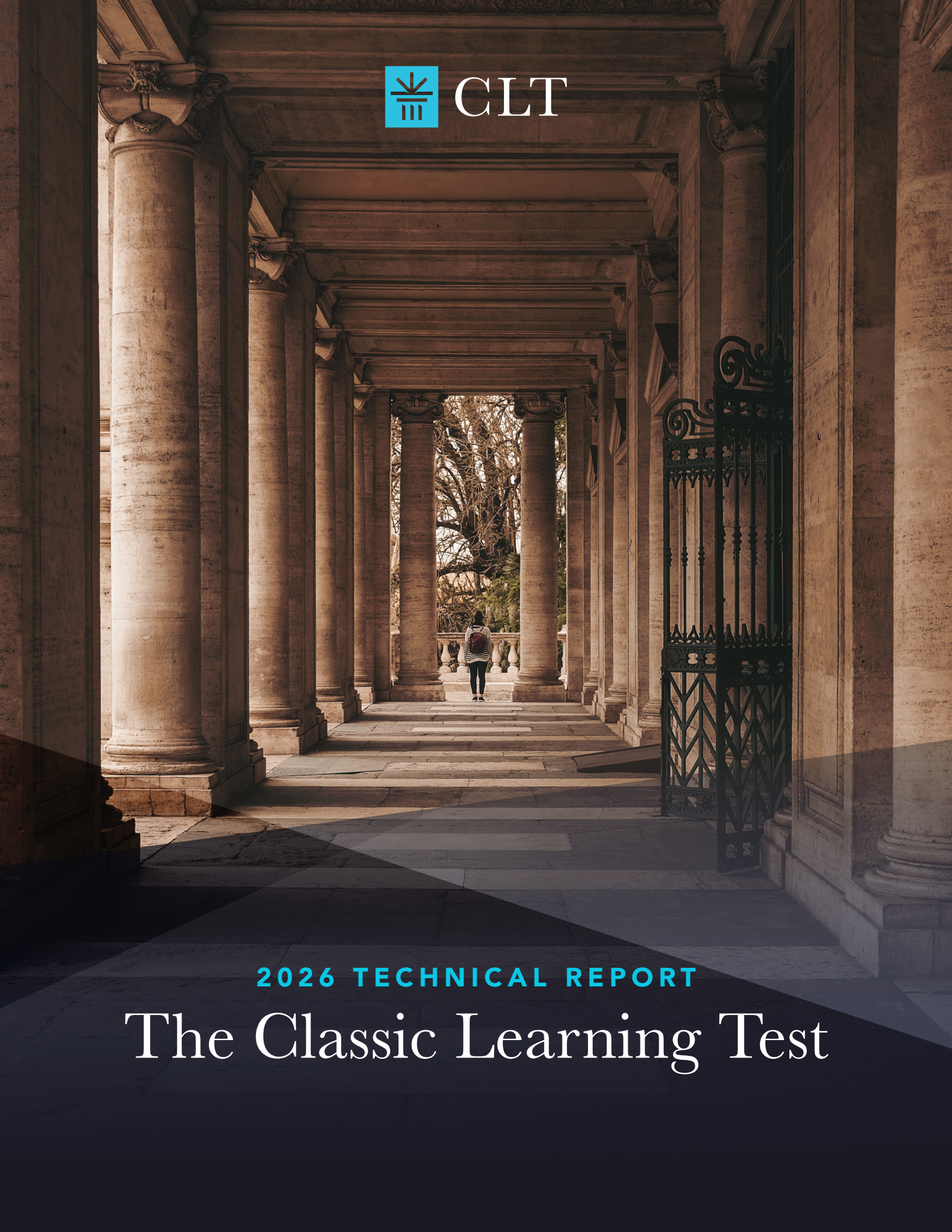


CLT

2026 TECHNICAL REPORT

The Classic Learning Test

CLT

2026 TECHNICAL REPORT

The Classic Learning Test

CONTRIBUTORS

Report Lead ..... Noah Tyler

Psychometricians ..... Eren Asena
Research/Statistical Analyst, CLT; MSc, Methodology & Statistics, University of Amsterdam

Technical Advisory Committee: Dr. Garron Gianopulos
Psychometric Consultant, Ph.D., Curriculum and Instruction: Educational Measurement and Research, University of South Florida

Dr. Hong Jiao
Psychometric Consultant, Ph.D., Measurement, Statistics, and Evaluation, Florida State University

Sylvia Tidwell Scheuring
Founder, President, and Chief Learning Officer of Arroki, Inc. B.S. Physics from the University of Arizona. PhD, Research in Educational Measurement & Statistics (ABD) University of Kansas

Writers & Reviewers ..... Tracy Gardner, Ph.D.
Psychometric Consultant (as of February, 2026) Research Methodology, University of Pittsburgh

Livvy Beaver

M. William Veitkus,
M.A.C.E., Hillsdale College

Graphic Designer ..... Meg Pilcher

Letter from the CEO

Standardized testing is often treated as something routine and procedural, but in reality, it exerts enormous influence over the direction of education. Tests do not simply measure what students have learned. They shape what schools prioritize, what teachers emphasize, and ultimately what students are asked to spend their time thinking about. In that sense, assessments do not merely reflect the curriculum. They help drive it.

That is why the design and administration of assessments carry such serious responsibility. In today's educational landscape, test scores are used to make high-stakes decisions for students, institutions, and policymakers. If those scores are going to carry that kind of weight, they must be interpretable, comparable, and supported by strong evidence. An assessment must demonstrate not only that its content is meaningful, but that its scores are valid, reliable, and fair across populations and administrations. When those conditions are met, testing can serve a legitimate and important role: recognizing achievement and helping connect students with opportunities that correspond to their preparation and potential.

At CLT, we approach assessment development with a commitment to both technical rigor and operational integrity. Our tests are built according to established psychometric principles, including clear construct alignment, deliberate item development and review, and continuous statistical evaluation. Reliability is examined through multiple indicators, and validity evidence is gathered to support the interpretation and use of scores. These are not box-checking exercises. They are essential to ensuring that a score has real meaning and can be used with confidence.

Test security is no less important. A score is only as trustworthy as the conditions under which it was earned. If results are shaped by item exposure, compromised administration conditions, or misconduct rather than actual student ability, the score itself begins to lose meaning. For that reason, CLT takes test security seriously, employing a comprehensive approach that includes controlled item exposure, clear proctoring protocols, data forensics, and ongoing monitoring for irregularities. These measures are necessary not simply to protect test content, but to preserve the fairness and interpretability of every score we report.

At its best, assessment serves a larger educational purpose. It can direct attention toward what is enduring, reward careful reasoning, and reinforce the kinds of intellectual habits that matter most. In that sense, assessment is not only about measurement. It is also about orientation. It reflects what a system of education believes is worth knowing, worth pursuing, and worth passing on.

CLT's work is guided by three core commitments.

Anchored, in the belief that assessment should remain connected to enduring intellectual traditions and foundational ideas.

Passionate, in our conviction that this work matters, and that quality, continuous improvement, and careful stewardship are all necessary if assessment is to be done well.

Humane, in our belief that testing should respect the dignity of the student and contribute, however modestly, to the formation of the whole person.

This Technical Report provides detailed documentation of the methodologies, analyses, and procedures that support the CLT. It is intended to offer transparency into how scores are developed, validated, and protected, and to support confidence in their appropriate interpretation and use. In doing so, it reflects our commitment to assessments that are secure, meaningful, and aligned with a richer vision of education and human flourishing.

Jeremy Tate

Jeremy Tate,
Founder and CEO of CLT

Executive Summary

The Classic Learning Test (CLT) is a standardized assessment designed to evaluate the extent to which students have developed the intellectual formation, expressed through literacy, numeracy, and reasoning, that supports success across postsecondary pathways, including entry-level, credit-bearing college coursework. CLT scores provide a clear, interpretable, and empirically supported signal to inform admissions, placement, and related decision-making processes.

CLT is intended for use with high school students across diverse educational settings, including traditional public schools, private schools, homeschools, and microschools. While broadly accessible as a measure of foundational academic skills, the assessment is particularly well-aligned to students whose preparation has emphasized engagement with complex, knowledge-rich texts, analytical reading, and the development of disciplined reasoning. This alignment strengthens the interpretability of scores for their intended uses.

The design of CLT is grounded in a coherent framework of learning and human development. Consistent with longstanding educational traditions, students first acquire essential intellectual skills such as reading, writing, and arithmetic, which enable them to pursue knowledge across core domains of human inquiry, including mathematics and the natural sciences, as well as history, philosophy, politics, and theology. CLT reflects this framework by assessing the literacy and reasoning capacities required to engage meaningfully with complex ideas across these domains. In doing so, the assessment captures the broader intellectual habits that support academic success and lifelong learning.

The validity argument for CLT is structured in accordance with professional standards for educational and psychological testing and reflects a chain of evidence linking assessment design, measurement, and score use. The assessment is designed to measure core constructs of verbal reasoning, writing, and quantitative reasoning, and to support inferences about students' intellectual formation and preparedness for postsecondary engagement. These inferences are supported by multiple sources of evidence, including content alignment, reliability analyses, concordance with established assessments, classification consistency, and fairness analyses.

CLT operates within a clearly articulated theory of action in which assessment design promotes disciplined engagement with meaningful ideas, leading to the development and demonstration of intellectual formation. This progression supports valid score interpretations and informs decisions related to admissions, placement, and postsecondary pathways. The assessment is therefore designed not only to measure outcomes, but also to reinforce alignment between curriculum, instruction, and the expectations of postsecondary environments.

CLT scores are used to support a range of postsecondary pathways, including but not limited to college enrollment, workforce preparation, technical programs, and alternative credentialing opportunities. By providing a valid and reliable measure of students' academic preparation and reasoning capacity, CLT contributes to improved placement decisions, stronger alignment between K–12 and postsecondary systems, and more effective student pathways.

Consistent with best practices in assessment and credentialing, CLT is supported by an ongoing program of research, validation, and continuous improvement. This includes regular evaluation of benchmarks, monitoring of population characteristics, continued study of relationships between scores and postsecondary outcomes, and systematic review of test design and content. These practices reflect a commitment to maintaining a high-quality, psychometrically sound assessment system and to supporting the appropriate interpretation and use of scores over time.

Taken together, CLT represents a coherent, evidence-based assessment system that integrates rigorous measurement with a clearly defined framework of learning. By aligning assessment design, score interpretation, and intended use, CLT provides institutions and students with a trustworthy tool for navigating postsecondary pathways and supports the broader aims of academic success, lifelong learning, and human flourishing.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 What is the CLT?

Classic Learning Initiatives (CLI) launched in December 2015 as an alternative to the College Board and ACT Inc. As of May 2023, more than five hundred thousand CLT assessments have been administered in homes and schools across the United States,[1] and over three hundred and fifty colleges and universities have adopted it as an admissions test.[2]

This vision is reflected in the design of the assessment itself. The CLT is grounded in a coherent Theory of Action that connects assessment design, student engagement, and student outcomes. At its core is a validity pathway for intellectual formation: assessment design leads to disciplined engagement with meaningful ideas, which cultivates intellectual formation and supports postsecondary success and human flourishing. In this way, the assessment is intended to measure academic skills and to reflect and reinforce the habits of mind and virtues that inspire students to pursue truth, goodness, and beauty throughout their lives.

The CLT is a different kind of standardized college entrance exam. It aims to dramatically enrich students' test-taking experience and to motivate positive change in assessment and education. The CLT is built on the idea that the purpose of education is to make us more human. Students must grapple with ideas that enable them to engage with profound truth, weigh evidence, understand different perspectives, and ultimately build a foundation that will serve them for the rest of their lives.

Frederick Douglass said, "Education means emancipation. It means light and liberty. It means the uplifting of the soul of man into the glorious light of truth, the light by which men can only be made free."

> **"Education means emancipation. It means light and liberty. It means the uplifting of the soul of man into the glorious light of truth, the light by which men can only be made free."**
>
> *Frederick Douglass*

1 The CLT suite of assessments is comprised of: the CLT, a college entrance exam; the CLT10, a preparatory exam for the CLT offered to 9th and 10th graders; the CLT8, an end-of-grade assessment tool designed for 8th-grade students as they prepare to enter high school; and the pilot CLT3-6 administered in the spring of 2023.

2 The full list of colleges which have adopted the CLT as an admissions exam is provided at https://www.cltexam.com/colleges.

It means light and liberty. It means the uplifting of the soul of man into the glorious light of truth, the light by which men can only be made free."

The CLT serves the needs of educators, students, and parents. Students take a shorter exam, either in school or at home with our remote proctoring services. Tests taken in school can be taken either online or in paper form (according to the school's preference). Testers and administrators access the exam's analytics through their online CLT accounts, and testers can send their scores to colleges for free. Furthermore, not only does the CLT challenge students, it also sets them apart from their peers in college applications.

1.2 Improving Students' Test-Taking Experience

For students, the CLT is refreshingly user-friendly and modern. It was designed with the goal of providing the best possible test-taking experience, and includes the following features:

- Online platform accessible via students' own desktops, laptops, or tablets
- Remotely Proctored exams are available for students testing at home.
- Paper tests for in-school testers
- Shorter test-taking time (120 minutes, not including 30 minute optional essay)
- Scores released the Wednesday after the exam for in-school testers and the third Wednesday after the exam for at-home testers
- In-depth Student Analytics

TEST MODES

The CLT is primarily administered online, though a paper version is available for in-school testing. The online platform is more natural for contemporary students than a pencil and paper format, and reduces the risk of confusion and unnecessary mistakes. Students can select and change their answers with one click, without having to fill in Scantron bubbles, take time to erase, or risk entering multiple answers.

Students testing online take the test on their own devices. Using an unfamiliar device for a high-stakes test can lead to a more frustrating test-taking experience, as every device has its own subtle differences; allowing students to use a device they are already familiar with reduces the possibility that the device itself will impair the student's ability to perform.

In the spring of 2020, the CLT launched a new test mode for students testing from home. The Remotely Proctored CLT is typically offered twelve times per year, and allows students to take the exam from their home. The remotely proctored exam is auto-timed, and incorporates screen-share and video recording technology, to ensure test integrity without requiring an in-person proctor.

PREDICTABLE FORMAT

The CLT is designed for simplicity and balance. Each of the three sections has forty (40) questions. Each Verbal Reasoning and Grammar/Writing section has four (4) reading passages, and each passage has ten (10) questions. Knowing what to expect frees students from anxieties that can come from an irregular test design.

The test aesthetic is clean and free from distraction. It uses a white background and a readable serif font, and the reading questions line up side by side with the passage.

STRAIGHTFORWARD SCORING

Every CLT has 120 scored questions for a total of 120 possible points; there is no penalty for incorrect answers. The 120-point scale allows the test to be divided into three equally valuable sections with 40 questions each. The total score that the student receives on the CLT closely approximates the number of test questions that the student answered correctly across all three sections. (In cases where an administered test is slightly less or more difficult than expected, statistical techniques are used to equate tests, ensuring that each test is of equal difficulty and thus that scores are genuinely equivalent.)

SHORT TEST—FAST RESULTS

The CLT is 120 minutes long, or two hours (not including the 30 minute optional essay). The CLT was designed to be shorter than comparable tests in order to take as little as possible away from instruction time. Moreover, any added information gathered by day-long or multi-day assessment regimes is of questionable value, due to evidence that many students' scores can be negatively affected by fatigue.

In-school testers that take the exam online can access their scores the within a week of examination.. Students who test using a paper-based test receive scores once the tests are scanned and processed, within 30 days of receipt of returned answer sheets. Testers who take the remotely proctored exam receive their scores within three weeks of exam administration.

IN-DEPTH ANALYTICS

As an online preparatory exam, CLT scores and analytics can be used to assess the students' readiness to begin college.

CLT analytics reports are straightforward and easy to interpret. They indicate performance on the exam across multiple academic domains and subdomains, as well as comparisons to past test performance.

Student-level analytics are available to all students, whether they took the test from home or in school. From the student portal, testers can access definitions of each subdomain, sample questions, and lists of the main skills being assessed.

School- and class-level analytics, as well as individual analytics reports, are available for school administrators and teachers to view once their school has administered a test. Teachers and administrators can use the analytics and related documents to understand individual student performance and aptitude.

1.3 Motivating Positive Change in Assessment and Education

The CLT aims to change the landscape of assessment, and education generally, by providing a rigorous, intellectually rich exam. CLT exams assess both aptitude and achievement, feature rich reading passages, and support strong educational choices.

APTITUDE AND ACHIEVEMENT

Students must draw upon the education they have received in order to demonstrate what they have learned. Achievement within a domain of knowledge is one key purpose of assessment, and a principle focus for the CLT. Students preparing for the CLT, and administrators reviewing analytics, want to know that their plan of content formation will put them on the right track to perform well on the exam. The domains and subdomains provide the basic content framework of the exam.

The CLT aims to assess not only students' achievement, but also their aptitude. Students at this stage in their education are discovering their innate intellectual potential. CLT measures skills students develop through a variety of education types, such as their ability to communicate clearly, to read complex prose, to understand metaphors, to think logically, and to solve puzzles. Some students have natural talent in one or more of these areas, and the CLT can help identify those aptitudes.

Because the CLT is both an achievement and aptitude test, students are provided a window into their own unique set of intellectual strengths, while also receiving the tools through CLT analytics to make incremental improvements in their less developed areas.

RICH READING PASSAGES

In the CLT Verbal Reasoning and Grammar/Writing sections, students engage works from the greatest minds in the history of the liberal arts tradition. The test draws on literary, philosophical, and scientific passages from a wide variety of thinkers, such as St. Augustine, Dante, Sir Isaac Newton, Charlotte Brontë, W. E. B. Du Bois, and many more. These sources are both secular and religious, contemporary and historical. They require students to analyze texts, comprehend great ideas, and engage with issues that affect the world at large.

The CLT's distribution of subject categories in passages is as follows. On every test, out of eight reading passages, two (25%) are in Philosophy/Religion; one (12.5%) is drawn from Literature; two (25%) are in Science; one (12.5%) is an excerpt from Historical/Founding Documents; one (12.5%) is a Historical Profile; and one (12.5%) is drawn from Modern/Influential Thinkers.

DISTRIBUTION OF SUBJECT CATEGORIES ACROSS CLT PASSAGES

PASSAGE TYPE	NUMBER OF PASSAGES PER TEST	EXAMPLES
Modern/Influential Thinkers	12.5% (1 passage)	*A World Split Apart* by Aleksandr Solzhenitsyn "Address to the Nation on the State of the U.S. Economy" by John F. Kennedy
Historical Profile	12.5% (1 passage)	*The Heart of a Woman* by Maya Angelou "Personal and Literary Character of Cicero" by John Henry Newman
Historical/Founding Documents	12.5% (1 passage)	"Federalist No. 37" by James Madison *Politics* by Aristotle
Literature	12.5% (1 passage)	*Emma* by Jane Austen *Crime and Punishment* by Fyodor Dostoevsky
Science	25.0% (2 passages)	*On the Motion of the Heart and Blood in Animals* by William Harvey *Insectivorous Plants* by Charles Darwin
Philosophy/Religion	25.0% (2 passages)	"Of the Origin of Ideas" by David Hume "A Farewell Sermon" by Jonathan Edwards

1.4 CLT in Context

CLT has deep relationships with secondary schools, institutions of higher learning, think tanks, education policy organizations, philanthropists, and lawmakers that are passionate about meaningful education and the liberal arts. By linking arms with these individuals and organizations, CLT seeks both support and counsel in its mission to provide unmatched assessments that reflect and strengthen a holistic education, whether public, private, charter, or classical. Our core values of remaining Anchored, Passionate, and Humane are invigorated and preserved by these vital relationships.

The CLT Board of Academic Advisors is composed of prominent scholars, thought leaders, and visionaries in education who advise and advocate for CLT, as well as provide expert guidance.

In addition to the distinguished list of educators in colleges and universities and in private, parochial, homeschool, and charter schools, the board has executive leaders from a variety of mission-aligned organizations. These include:

- Classical Academic Press
- The Circe Institute
- Classical Conversations
- The Society for Classical Learning
- Hillsdale College K-12 Education
- Memoria Press
- The Association of Classical Christian Schools
- The American Council of Trustees and Alumni
- The Heritage Foundation
- The Institute for Catholic Liberal Education

A complete list of CLT board members can be found on our website.

1.5 About the CLT Technical Report

This technical report is a guide explaining the details of how the CLT exam works. Chapters 1-5 describe the design and administration of the CLT, and Chapters 6-11 explain and analyze the test's metrics.

Chapter 2 presents the content of the test itself, including sample questions, the author bank, and information on how test questions are organized by difficulty level. Chapter 3 outlines the steps CLT takes to develop, edit, and prepare each test for administration. Chapters 4 and 5 explain how the CLT is administered and describe the measures taken to ensure the test's security and fairness.

Chapter 6 provides information on how CLT scores are reported to students, administrators, and colleges. Chapter 7 provides background on Classical Item Analysis. Chapter 8 explains how tests are scaled using Item Response Theory. Chapters 9 and 10 quantify the test's reliability and validity, respectively. Chapter 11 presents norming evidence, including CLT/SAT concordance charts.

2.
STANDARDS AND CONTENT COVERAGE

2.1 Overview *(of the CLT Assessments, Skills Measured, and Design)*

The Classic Learning Test (CLT) was created in the context of a national movement to renew the foundations of great education. "Classic" here simply means an assessment that reflects tried and true ideas rather than contemporary experiments.

Although the CLT is open to all test-takers, the intended test-taking population is all 11th and 12th grade students in the U.S. and internationally. The target population of CLT test-takers consists of students in non-district schools: homeschool, private, parochial, and charter schools. The CLT is, however, well-suited for any student aspiring to high standards of literacy and numeracy.

The liberal arts education model trains students in language arts and mathematics as a path "to make the acquisition of all later studies more simple and effective."[1] Clark and Jain (2013) write, "Recovering the primacy of both the language arts and the mathematical arts is a pivotal piece of this paradigm. Together they train the student not just in what to think but in how to think."[2] In this way, the CLT exam draws on enduring concepts accessible to students from a variety of educational backgrounds. These include perennial questions about human nature and the physical world; lessons from history; and universal mathematical concepts.

The construct to be measured on the CLT exam, which underlies the CLT score, is a measure of a student's grammatical, logical, rhetorical, quantitative, and critical-thinking skills expected at the college level.

The purpose of the CLT exam is to focus on foundational intellectual skills such as clear reasoning and critical thinking, while tapping into the deep intellectual tradition of the classics. This approach to testing is aimed to measure not just students' academic achievements, but also their aptitude—to allow students to demonstrate their intellectual capabilities, regardless of their prior academic training.

Each CLT exam consists of three mandatory sections—Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning—as well as an optional Essay.

1 Clark, Kevin and Ravi Jain. *The Liberal Arts Tradition: A Philosophy of Christian Classical Education*. Classical Academic Press, 2013.
2 Ibid.

OVERVIEW OF CLT FORMAT

Section	Time allotted	Number of Questions
Verbal Reasoning	40 minutes	40
Grammar/Writing	35 minutes	40
Quantitative Reasoning	45 minutes	40
Totals:	2 hours*	120

**2 hours and 30 minutes with the optional essay*

These are similar to the sections in the SAT™ and are recognizable to students familiar with other standardized tests, but the content of the test is distinct from other standardized tests in two main ways.

First, CLT's two English sections primarily use selections from time-tested authors who have shaped history, literature, and philosophy in foundational ways through the centuries. The CLT thus provides an opportunity for students to interact with important thinkers whose voices have made a profound difference in the world of ideas.

Second, the Quantitative Reasoning section assesses students' ability to solve problems and to think in a logical and creative manner. The test focuses on assessing mathematical reasoning capacity in addition to testing specific mathematical skills or knowledge.

DIFFICULTY LEVELS

Reading passages in the Verbal Reasoning and Grammar/Writing sections are calibrated to fit narrowly within a consistent difficulty level. The test developers use a variety of tools, including a passage calibration software with grade-level ratings, to help analyze the difficulty level of each passage and ensure it falls within an appropriate range.

Difficulty levels of questions are scored on a scale of 1 through 5: each section of the test contains eight questions at each difficulty level, for a total of twenty-four questions at each difficulty level across the exam. In the Verbal Reasoning and Grammar/Writing section, difficulty levels are distributed evenly throughout each passage. Each passage, for which there are ten questions, has two questions of each difficulty level. In the Quantitative Reasoning section, questions increase in difficulty as they progress.

Level 1 questions are the least difficult, and require straightforward reasoning, basic logic, and a minimal number of steps to answer. Level 5 questions are the most difficult, and require more complex reasoning, higher-level thinking, and the ability to synthesize difficult concepts.

2.2 Author Bank

Education is not just about results. At CLT, we believe standardized testing provides students an invaluable opportunity to engage with the texts and authors that have shaped history and culture. Two thirds of CLT reading and writing passages are drawn from the list of authors below.

The CLT's focus on the Western and classical traditions presents students with ideas, themes, and arguments they will encounter for the rest of their lives. The men and women who have contributed to this intellectual canon come from all times and places, races and religions, classes and cultures.

CHAPTER TWO

ANCIENTS

The *Epic of Gilgamesh*, 18th c. BC?

Homer, 9th c. BC?

Hesiod, 8th c. BC?

Æsop, 621-565 BC

Confucius, 551-479 BC

Æschylus, 525-455 BC

Sophocles, 496-406 BC

Herodotus, 484-425 BC

Euripides, 480-406 BC

Thucydides, 460-400 BC

Hippocrates, 460-370 BC

Plato, 428-347 BC

Aristotle, 382-322 BC

Euclid, 4th-3rd c. BC

Archimedes, 287-212 BC

Terence, 195-159 BC

Cicero, 106-43 BC

Julius Cæsar, 100-44 BC

Lucretius, 99-55 BC

Virgil, 70-19 BC

Livy, 59 BC-AD 17

Ovid, 43 BC-AD 17

Seneca the Younger, 4 BC-AD 55

Josephus, 37-100

Plutarch, 46-120

Epictetus, 55-135

Tacitus, 56-120

Tertullian, 160-220

Origen, 184-253

St. Athanasius, 297-373

St. Gregory of Nyssa, 335-395

St. Jerome, 342-420

St. Augustine of Hippo, 354-430

MEDIEVALS

Boethius, 477-524

St. Benedict, 480-547

Procopius, 500-570

St. Gregory the Great, 540-604

St. Bede the Venerable, 673-735

Beowulf, 9th c.?

The Thousand and One Nights, 9th c.

Avicenna, 980-1037

St. Anselm of Canterbury, 1034-1109

Peter Abælard, 1079-1142

St. Bernard of Clairvaux, 1090-1153

Hugh of St. Victor, 1096-1141

St. Hildegard of Bingen, 1098-1179

Héloïse d'Argenteuil, 1100-1164

Averroës, 1126-1198

Moses Maimonides, 1138-1204

Marie de France, 1160-1215

The *Nibelungenlied*, c. 1200

Magna Carta, 1215

St. Thomas Aquinas, 1225-1274

The Saga of Erik the Red, 13th c.

Dante Alighieri, 1265-1321

Giovanni Boccaccio, 1313-1375

John Wycliffe, 1328-1384

Geoffrey Chaucer, 1343-1400

Julian of Norwich, 1343-1420

St. Catherine of Siena, 1347-1380

Christine de Pizan, 1364-1430

The *Pearl* Poet, 14th c.

St. Thomas à Kempis, 1380-1471

Thomas Malory, 1415-1471

EARLY MODERNS

Desiderius Erasmus, 1466-1536

Niccolò Machiavelli, 1469-1527

Nicolaus Copernicus, 1473-1543

St. Thomas More, 1478-1535

Martin Luther, 1483-1546

Bartolomé de las Casas, 1484-1566

John Calvin, 1509-1564

St. Teresa of Ávila, 1515-1582

Michel de Montaigne, 1533-1592

Francis Bacon, 1561-1626

William Shakespeare, 1564-1616

Galileo Galilei, 1564-1642

John Donne, 1572-1631

Thomas Hobbes, 1588-1679

René Descartes, 1598-1650

John Milton, 1608-1674

Blaise Pascal, 1623-1662

Margaret Cavendish, 1623-1673

Robert Boyle, 1627-1691

John Bunyan, 1628-1688

John Locke, 1632-1704

Isaac Newton, 1642-1727

Gottfried Leibniz, 1646-1716

Charles Montesquieu, 1689-1755

Voltaire, 1694-1778

Jonathan Edwards, 1703-1758

Benjamin Franklin, 1706-1790

David Hume, 1711-1776

Jean-Jacques Rousseau, 1712-1778

Adam Smith, 1723-1790

Immanuel Kant, 1724-1804

Edward Gibbon, 1737-1794

Antoine Lavoisier, 1743-1794

Thomas Jefferson, 1743-1826

Olaudah Equiano, 1745-1797

Johann Wolfgang von Goethe, 1749-1832

James Madison, 1751-1836

Mary Wollstonecraft, 1759-1797

Georg W. F. Hegel, 1770-1831

LATE MODERNS

Jane Austen, 1775-1817

Jakob & Wilhelm Grimm, 1785-1863 & 1786-1859

Mary Shelley, 1797-1851

Sojourner Truth, 1797-1883

St. John Henry Newman, 1801-1890

Alexis de Tocqueville, 1805-1859

Hans Christian Andersen, 1805-1875

John Stuart Mill, 1806-1873

Edgar Allan Poe, 1809-1849

Charles Darwin, 1809-1882

Charles Dickens, 1812-1870

Søren Kierkegaard, 1813-1855

Charlotte Brontë, 1816-1855

Henry David Thoreau, 1817-1862

Karl Marx, 1818-1883

Frederick Douglass, 1818-1895

George Eliot, 1819-1880

Herman Melville, 1819-1891

Susan B. Anthony, 1820-1906

Fyodor Dostoevsky, 1821-1881

Gregor Mendel, 1822-1884

Louis Pasteur, 1822-1895

Leo Tolstoy, 1828-1910

Mark Twain, 1835-1910

Friedrich Nietzsche, 1844-1900

Oscar Wilde, 1854-1900

Sigmund Freud, 1856-1939

Anna Julia Cooper, 1858-1964

Anton Chekov, 1860-1904

Alfred North Whitehead, 1861-1947

Ida B. Wells, 1862-1931

W. E. B. Du Bois, 1868-1963

Mahatma Gandhi, 1869-1948

Willa Cather, 1873-1947

G. K. Chesterton, 1874-1936

Albert Einstein, 1879-1955

Virginia Woolf, 1882-1941

John Maynard Keynes, 1882-1946

Franz Kafka, 1883-1924

Ludwig Wittgenstein, 1889-1951

Zora Neale Hurston, 1891-1960

J. R. R. Tolkien, 1892-1973

Dorothy Sayers, 1893-1957

F. Scott Fitzgerald, 1896-1940

C. S. Lewis, 1898-1963

Ernest Hemingway, 1899-1961

Jorge Luis Borges, 1899-1986

Friedrich Hayek, 1899-1992

Langston Hughes, 1901-1967

John Steinbeck, 1902-1968

George Orwell, 1903-1950

Hannah Arendt, 1906-1975

Albert Camus, 1913-1960

Aleksandr Solzhenitsyn, 1918-2008

James Baldwin, 1924-1987

Flannery O'Connor, 1925-1964

Elie Wiesel, 1928-2016

Martin Luther King, Jr., 1929-1968

Toni Morrison, 1931-2019

2.3 Verbal Reasoning Test

The Verbal Reasoning section tests a student's ability to understand and analyze a text. Students are asked to interact with a variety of texts in different subject areas, described in the subsection "Passage Types", and are tested on their ability to comprehend the text and synthesize ideas within that text. They must be able to understand concepts such as how different phrases and words are used in context, the author's purpose in a particular section or in the passage overall, how a text is structured, and what could be reasonably inferred based on the information in the text. This section contains 40 questions and the standard administration time is 40 minutes.

QUESTION TYPES

Each passage has ten questions. They are not ordered by level of difficulty. Each passage has two questions of each difficulty level. Below is the high-level test blueprint along with a description of each question type within the Verbal Reasoning section.

Comprehension (27 questions)

- Passage as a Whole: These question types measure the student's ability to synthesize information from an entire passage in order to understand its framework and main ideas. (8 questions)
- Passage Details: These question types measure the student's ability to understand key facts and concepts discussed in a passage. (11 questions)
- Passage Relationships: These analogy questions measure the student's ability to recognize important connections between different parts of a passage. (8 questions)

Note: Analogies require students to be able to connect high-level concepts within a passage and to make connections between ideas and terms in a passage. CLT's analogies refer to concepts within a passage and use terms students are likely to know already, rather than relying on difficult vocabulary to challenge students.

Analysis (13 questions)

- Textual Analysis: These question types measure the student's ability to make inferences from information in a passage and to understand a character, a narrator, or a writer's point of view. (8 questions)
- Interpretation of Evidence: These question types measure the student's ability to understand how verbal and quantitative evidence are used in a passage. (5 questions) One of the Interpretation of Evidence questions always refers to a figure accompanying the second passage of the four, which is always the Science passage.

Passage Types

Each Verbal Reasoning section consists of four passages: three full passages and one passage composed of two shorter excerpts presented together. Each Verbal Reasoning passage fits narrowly within a word count range of 500-650 words. The total word count for all passages within the Verbal Reasoning Section must be between 2,200-2,400, for an average of 2,300 words total.

The majority of the material in the Verbal Reasoning section is drawn from passages in the Western intellectual tradition (see the Author Bank on pages 13 to 15). The passages fall into four categories, which are consistent, including in order, across each exam:

- Literature: The passages in the Literature category are drawn from classic and modern literary prose. Authors include those whose stories, style, and ideas have contributed significantly to Western culture.

- Science: The passages in the Science category are from articles, essays, and other works exploring various disciplines such as genetics, astronomy, physics, biology, and chemistry. When relevant, these passages may touch on the ethical, moral, or societal implications of the work. Each science passage in the Verbal Reasoning section will be accompanied by a graphic, such as a chart or table.
- Philosophy/Religion: The passages in the Philosophy/Religion category are from contemporary or classic sources, and are concerned with issues of truth, reasoning, ethics, and more. They are drawn from a variety of perspectives and periods.
- Historical/Founding Documents: The paired passages in the Historical/Founding Documents category are two brief selections that present perspectives on a topic. The first is a historical document, often drawn from ancient sources. The second is a passage from a writer or time period significant to U.S. history.

For anything to be read or communicated, some common context is assumed. For example, a math question involving a six sided die does not explain what a die is. Tests with the most universally accessible design still do not remove all such questions. Like other fairly designed tests of verbal reasoning constructs similar to it, the CLT neither tests knowledge about specific information from outside of its given texts, nor does it avoid asking questions assuming some shared background information.

Further, the CLT tends to include passages of relevance, meaning, and weight: passages that have explicit societal and personal implications, that give historical perspectives and references, and that have had an influence on human history. The CLT does not test "specific, communally shared information", what E. D. Hirsch calls "acculturation", but neither is it shy from the fact that a wide understanding of literacy lies behind understanding a text with *any degree* of meaning, relevance, or weight. Hirsch (1987) describes this wide sense of literacy:

> **"It is the background information, stored in their minds, that enables them to take up a newspaper and read it with an adequate level of comprehension, getting the point, grasping the implications, relating what they read to the unstated context which alone gives meaning to what they read."**
>
> *E.D. Hirsch*

"What [Professor Chall] calls world knowledge I call cultural literacy, namely, the network of information that all competent readers possess. It is the background information, stored in their minds, that enables them to take up a newspaper and read it with an adequate level of comprehension, getting the point, grasping the implications, relating what they read to the unstated context which alone gives meaning to what they read."[3]

The CLT both seeks a universally accessible test design and recognizes that a student with a wider context of literacy will be more comprehending of and conversant with CLT texts.

3 Hirsch, E.D. *Cultural Literacy: What Every American Needs to Know.* Houghton Mifflin Company, 1987.

SAMPLE QUESTIONS

Below is one sample question for each subdomain in the Verbal Reasoning section.

Passage as a Whole

Overall, the passage can be best described as

A) a subtle exploration of the rivalry between two colleagues.
B) a whimsical tale of a fantastic beast.
C) a cogent story about an attempt to seek out novelty.
D) a meandering account of the sale of a crocodile.

Passage adapted from Fyodor Dostoevsky's "The Crocodile," 1865.

Passage Details

According to the passage, what is a hallmark of Mr. Pecksniff's character?

A) Suspicion of conventional morality
B) Affection for eloquent language
C) Fear of the unknown
D) Disinterest in the lives of his children

Passage adapted from Charles Dickens' Life and Adventures of Martin Chuzzlewit, *1844.*

Passage Relationships (Analogies)

medicine : body ::

A) exercise : spirit
B) philosophy : soul
C) politics : philosophy
D) love : friends

Passage adapted from Plutarch's "On Education" in Moralia, first century AD.

Textual Analysis

In Passage 1, Philosophy indicates she believes Socrates was put to death primarily because

A) his philosophy was ill-formed and only partial.
B) he traveled to a distant, violent land filled with barbaric tribes.
C) his allies, Anaxagoras and Zeno, did not support him.
D) he lived an upright, ethical life in contrast to those around him.

Passage adapted from The Consolation of Philosophy *by Boethius, sixth century AD.*

Interpretation of Evidence

Which lines in the passage provide the best evidence in support of the answer to the previous question?

A) Paragraph 4, Sentence 1 ("And this . . . reality")
B) Paragraph 4, Sentence 2 ("The great . . . fertilize")
C) Paragraph 5, Sentence 2 ("But the . . . tendency")
D) Paragraph 6, Sentence 1 ("Consequently . . . study")

Passage adapted from Christopher Dawson's Religion and the Rise of Western Culture: The Classic Study of Medieval Civilization, *1950.*

2.4 Grammar/Writing Test

The Grammar/Writing section tests a student's ability to edit and improve a text. Students are asked to interact with a variety of texts in different subject areas, described in the subsection "Passage Types", and are tested on their ability to correct errors within that text and to improve its readability and flow. The section assesses students on their ability to use punctuation correctly, to convey points precisely and concisely, to make appropriate transitions, to choose the correct part of speech, to match verb tense, and to make other grammatically well-formed choices. This section contains 40 questions and the standard administration time is 35 minutes.

QUESTION TYPES

Each passage has ten questions which are not ordered by level of difficulty. Each passage has two questions of each difficulty level. Each question requires students to either correct an error or suggest an improvement to the passage. If no change is necessary, students can select the option "NO CHANGE."

Below is a high-level test blueprint along with a description of each question type within the Grammar & Writing section.

Grammar (20 questions)

- Agreement: These question types measure the student's ability to recognize how individual elements of a sentence correspond to or agree with one another. (10 questions)
- Punctuation and Sentence Structure: These question types measure the student's ability to understand how different elements of a sentence are linked by punctuation, and how to properly construct a sentence. (10 questions)

Writing (20 questions)

- Structure: These question types measure the student's ability to recognize how different parts of a passage, paragraph, and sentence relate to one another. "Structure" questions often propose a structural change in the question stem, and offer two answer choices supporting the change for different reasons, and two answer choices rejecting the change for different reasons. For this reason, it is the only Grammar/Writing question type where choice A might be something other than "NO CHANGE."(8 questions)
- Style: These question types measure the student's ability to understand a writer's tone and intent. (8 questions)
- Word Choice: These question types measure the student's ability to recognize how different words fit into different contexts. (4 questions)

PASSAGE TYPES

The majority of the material in the Grammar/Writing section is drawn from the Author Bank (as in the Verbal Reasoning section). Tests are calibrated so that each Grammar/Writing passage fits narrowly within a word count range of 460-565 words. The total must be between 2,000-2,200 words, for an average of 2,100 words total.

The passages used in the Grammar/Writing section fall into four categories that remain consistent, in order as well as category, across each exam:

- Philosophy/Religion: The passages in the Philosophy/Religion category are from contemporary or classic sources that reason about issues of truth, ethics, and what it means to be human. They are drawn from a variety of perspectives and periods.
- Historical Profile: The passages in the Historical Profile category consist of short biographical pieces on important historical figures (e.g. Alexander the Great, St. Joan of Arc, William Shakespeare, and Harriet Tubman).
- Science: The passages in the Science category are from articles, essays, and other works exploring various disciplines such as genetics, astronomy, physics, biology, and chemistry. When relevant, these passages may touch on the ethical, moral, or societal implications of the given work. Science passages in Grammar/Writing sections do not include a table or graph as they do in Verbal Reasoning sections.
- Modern/Influential Thinker: The passages in the Modern/Influential Thinker category are similar in scope to the Philosophy/Religion category, but are always drawn from more modern sources, and may offer perspectives on issues currently faced by society.

SAMPLE QUESTIONS

Below is one sample question for each subdomain in the Grammar/Writing section.

Agreement

caring decisions

A) NO CHANGE
B) caringly decisions
C) careful decisions
D) carefully decisions

Passage adapted from Hilaire Belloc's The French Revolution, *1911.*

Punctuation and Sentence Structure

in the National Government—in the Congress and in the States—to

A) NO CHANGE
B) in the National Government; in the Congress; and in the States—to
C) in the National Government, in the Congress and in the States to
D) in the National Government, in the Congress, and in the States to

Passage adapted from John F. Kennedy's "Address to the Nation on the State of the U.S. Economy," 1962.

Structure

The author wants to add a sentence to the end of this paragraph. Which option fits best in the passage?

A) Pell never solved the ancient problems of Diophantos, however.
B) By 1800, independent projects had listed the primes up to 1 million.
C) Unfortunately, most of these numbers were incorrect.
D) Pell would have been able to create two million primes had he had a computer.

Passage adapted from Martin H. Weissman's "Why prime numbers still fascinate mathematicians, 2,300 years later," 2018.`

Style

Of course, from the hearts of human beings, laws will not eliminate prejudice from them.

A) NO CHANGE
B) Of course, from human beings' hearts, prejudice will not be eliminated by human laws they create.
C) Of course laws will not eliminate prejudice from the hearts of human beings.
D) Laws of the hearts of human beings are not eliminated by prejudice, of course.

Passage adapted from Shirley Chisholm's "For the Equal Rights Amendment," 1970.

Word Choice

permeated

A) NO CHANGE
B) persisted
C) persecuted
D) persevered

Passage adapted from St. Teresa of Ávila's The Way of Perfection, *1583.*

2.5 Quantitative Reasoning Test

The Quantitative Reasoning section tests students' ability to think logically, use and manipulate symbols, and understand shapes. Students are asked to complete a variety of questions of various subtypes in order to assess their logical reasoning ability across different domains.

As one can gather from the question types described on the following pages, the Quantitative Reasoning section of the CLT tests algebra I and II and geometry, including coordinate plane geometry and trigonometry. The CLT intends to measure creative skills beyond those content specific, algorithmic ones, however— skills like numeracy, facility with numbers and the manipulation of expressions, fine-tuned mathematical intuitions, and creative approaches to unfamiliar problems. One might be surprised to see a question about odd and even numbers, for example, on a test intended for 11th and 12th grade students; but one might find these questions among the most difficult for some students, because they ask for a working understanding of or intuition about number theory.

Calculators are not allowed on the exam. Basic formulas are provided for each exam; formulas below are accessible by clicking the "Show Formulas" button on the left side of the page.

Area of a circle $= \pi r^2$, where r is the radius of the circle

Circumference of a circle $= 2\pi r$, where r is the radius of the circle

There are 360 degrees in a circle.

There are 2π radians in a circle.

Volume of a sphere $= \frac{4}{3}\pi r^3$, where r is the radius of the sphere

Surface area of a sphere $= 4\pi r^2$, where r is the radius of the sphere

Area of a rectangle $=$ length $\times$ width

Area of a triangle $= \frac{1}{2}$(base $\times$ height)

The sum of the measures of the interior angles of a triangle is 180°.

Pythagorean theorem (for a right triangle): If a, b, and c are the side lengths of the triangle, and c is the hypotenuse, then $a^2 + b^2 = c^2$.

Trigonometry:

$\sin\theta = \frac{\text{opposite}}{\text{hypotenuse}}$

$\cos\theta = \frac{\text{adjacent}}{\text{hypotenuse}}$

$\tan\theta = \frac{\text{opposite}}{\text{adjacent}}$

$\csc\theta = \frac{1}{\sin\theta}$

$\sec\theta = \frac{1}{\cos\theta}$

$\cot\theta = \frac{1}{\tan\theta}$

$\tan\theta = \frac{\sin\theta}{\cos\theta}$

$\sin^2\theta + \cos^2\theta = 1$

$30^\circ\text{–}60^\circ\text{–}90^\circ$ triangles have side lengths in a ratio of $1 : \sqrt{3} : 2$, corresponding to their opposite angle.

$45^\circ\text{–}45^\circ\text{–}90^\circ$ triangles have side lengths in a ratio of $1 : 1 : \sqrt{2}$, corresponding to their opposite angle.

QUESTION TYPES

In the Quantitative Reasoning section, questions are broken down into three main types:

Algebra I and II:

The 10 questions in the Algebra category include problems on properties of integers, substitution, sequences, systems of equations, quadratic equations, etc.

- Arithmetic and Operations: These question types measure the student's ability to use basic rules of arithmetic to simplify and draw conclusions about expressions, as well as the ability to recognize patterns.
- Algebraic Expressions and Equations: These question types measure the student's ability to simplify algebraic expressions—which, unlike the expressions in "Arithmetic and Operations" questions, usually include variables—solve equations and inequalities, and substitute variables into algebraic expressions.

Geometry:

The 14 questions in the Geometry category test a student's ability to analyze shapes and determine key pieces of information from what is given in a problem. Students may be tested on polygons, properties of parallel and perpendicular lines, coordinate geometry, and trigonometry. The CLT emphasizes intuitive use of geometric principles rather than memorization of formulas.

- Plane Geometry: These question types measure the student's ability to analyze two-dimensional shapes and to understand points, lines, figures, and functions in the (x,y)-coordinate plane.
- Properties of Shapes: These question types measure the student's ability to analyze circles, triangles, and other polygons and determine additional information about those shapes.
- Trigonometry: These question types measure the student's ability to use a right triangle's angle measurements and the ratios between its side lengths in order to deduce additional information. Advanced questions also look at a student's ability to understand and manipulate trigonometric identities, analyze trigonometric functions on the unit circle, and graph trigonometric functions.

Mathematical Reasoning:

The 16 questions in the Mathematical Reasoning category will most often be word problems that require students to apply logic and reasoning to given situations. Problems may include properties of integers, geometric shapes, ratios, or algebra. Some questions will ask students to draw conclusions based on a set of given conditions.

- Logic: These question types measure the student's ability to validly deduce a conclusion from given information.
- Word Problems: These question types measure the student's ability to use reasoning and logic to draw conclusions in real-life scenarios.

SAMPLE QUESTIONS

Below is one sample question for each subdomain in the Quantitative Reasoning section.

Arithmetic and Operations

The expression $2^7 + 2^7$ is equivalent to which of the following?

- A) 2^8
- B) 2^9
- C) 2^{14}
- D) 2^{49}

Algebraic Expressions and Equations

What are the x-coordinates of the points of intersection of the parabola $y = x^2 - 7$ and the line $y = x - 1$?

- A) $x = 1$, $x = \sqrt{7}$, and $x = -\sqrt{7}$
- B) $x = 1$ and $x = 3$
- C) $x = -2$ and $x = -3$
- D) $x = -2$ and $x = 3$

Plane Geometry

Line L is parallel to the line $2y - 3x = 7$. Which of the following is perpendicular to line L?

- A) $y = \frac{3}{2}x - 7$
- B) $y = -\frac{1}{6}x + 7$
- C) $y = -\frac{2}{3}x + 7$
- D) $y = \frac{3}{2}x - \frac{1}{7}$

Properties of Shapes

The perimeter of one face of a cube is 20 cm. What is the surface area of the cube?

- A) $25\,\text{cm}^2$
- B) $50\,\text{cm}^2$
- C) $150\,\text{cm}^2$
- D) $600\,\text{cm}^2$

Trigonometry

Which of the following is equivalent to the expression $\frac{\sin x \sec x}{\sin^2 x + \cos^2 x}$?

- A) $\sin x$
- B) $\cos x$
- C) $\tan x$
- D) $\sin x \cos x$

Logic

A student has invented the following rule for right triangles:

All right triangles have side lengths in the ratio of 3:4:5.

Which of the following is a counterexample that disproves the above statement?

- A) A triangle with side lengths 2, 3, and 4.
- B) A triangle with side lengths 5, 12, and 13.
- C) A triangle with side lengths 6, 8, and 10.
- D) A triangle with side lengths 7, 7, and 10.

Word Problems

At a gift store, candles are sold in packages of 4, chocolates are sold in packages of 10, and thank-you cards are sold in packages of 3. Miranda is putting together gift bags, each of which contains one candle, one chocolate, and one card. What is the smallest number of gift bags she can make such that she doesn't have any items left over?

- A) 20
- B) 30
- C) 60
- D) 120

2.6 Optional Essay

Testers who take the exam online in a school-proctored setting have the option of completing an unscored essay section. This essay gives students the opportunity to provide colleges with an example of their writing ability under a time limit. Students have 30 minutes to answer one prompt. Their written response is included with their test results when students send their scores to colleges.

Sample essay prompts are as follows:

SAMPLE ESSAY 1: Describe what you believe a community to be. What defines it? How large is it? What are its boundaries, and what determines who is inside and out of it? You can draw on contemporary, historical, or literary examples to support your claims.

SAMPLE ESSAY 2: The Stoic philosophers were deeply concerned by emotion and its tendency to overwhelm. Can emotion be a good thing? Is it a threat to reason, or can it aid reason? Provide examples from history or literature to support your claims.

SAMPLE ESSAY 3: Are there any situations in which censorship of works is appropriate? If so, explain in what context and why. If not, explain why not. Use examples to support your claims.

3. TEST DEVELOPMENT

Overview

The Test Development team of CLT writes and edits each test item according to a specific set of parameters. The Test Development and Operations teams work together in the test preparation process, following a schedule of development, review, and uploading, so that every test undergoes quality control and is ready on time.

3.1 Test Blueprint

The CLT test blueprint defines the structure and content specifications for each assessment form. It establishes the distribution of items across content domains, cognitive demands, and difficulty levels, ensuring that each form measures the intended constructs of verbal reasoning, grammar/writing, and quantitative reasoning.

The blueprint is designed to support alignment between assessment content and the knowledge and skills associated with postsecondary readiness. It also ensures consistency across forms, enabling comparability of scores across administrations.

The test blueprint serves as the foundational design specification for all operational forms and is reviewed periodically to ensure continued alignment with the constructs being measured and the intended uses of the assessment. Updates to the blueprint are informed by empirical item performance data, content coverage analyses, and evolving expectations related to postsecondary readiness. All blueprint revisions are documented and version-controlled to ensure consistency across administrations.

3.2 Selecting and Training Item Developers

CLT selects item developers based on subject-matter expertise and familiarity with the types of texts and reasoning skills emphasized in the assessment. Item writers are drawn from a range of academic and educational backgrounds and are expected to demonstrate strong content knowledge as well as the ability to develop questions aligned to complex, knowledge-rich passages.

Item developers are trained to align their work with the CLT blueprint and to produce items that reflect the assessment's emphasis on reasoning, clarity, and engagement with meaningful content.

All item writers undergo structured training that includes guidance on item design principles, alignment to the test blueprint, avoidance of construct-irrelevant variance, and adherence to CLT content and style standards. Training materials and expectations are reviewed periodically and updated as needed to ensure consistency and quality across item development cycles.

3.3 Form Assembly

CLT forms are assembled to ensure alignment with the test blueprint and to maintain consistency in content coverage, difficulty, and construct representation across administrations. Each test form is constructed using calibrated items that have been evaluated through statistical analysis. The assembly process ensures that items function together to produce reliable and interpretable scores.

Test Information Functions (TIFs) and Test Characteristic Curves (TCCs) are used to evaluate the overall functioning of each form and to confirm comparability across forms. TIFs graph the amount of information provided by the items on a form about student abilities as estimated by the Rasch model. TCCs provide a graphical representation of the relationship between ability as estimated by the Rasch model and expected raw scores on a form. TIFs and TCCs allow us to evaluate the degree to which different forms are parallel to each other in terms of difficulty and reliability.

AUTOMATED TEST ASSEMBLY (ATA)

When test development begins for the upcoming academic year, test forms are assembled as section modules that follow the content blueprint and certain statistical specifications. A module is a mini test form that consists of a single section. CLT uses automated test assembly (ATA) to construct modules that are parallel in content and statistical specifications. ATA is conducted using the *eatATA* package (Becker et al., 2021) in the R programming language (R Core Team, 2023). ATA involves computer algorithms that translate a set of constraints defined by psychometricians and content experts into mathematical optimization problems. Constraints related to the content of the tests come from the blueprints created by our test development team and include the number of items each form should include from each subject, domain, subdomain, passage type, and question type.

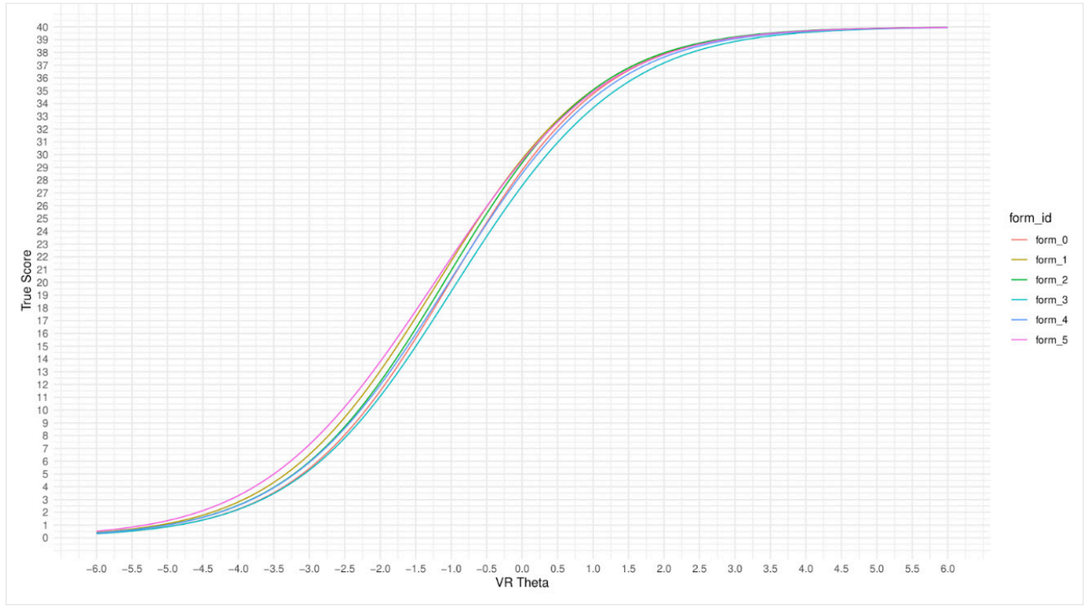
Then, statistical constraints are set to ensure that only high-quality items are included in the forms. High-quality items are those that can discriminate well between high ability and low ability students and do not show bias against a particular demographic subgroup such as a gender or an ethnicity. Item discrimination is measured by the point-biserial correlation between item scores-Rasch thetas and item scores-total scores, as explained in Chapter 8. To investigate bias, we analyze differential item functioning (DIF), which is explained in Chapter 10. Items that are flagged due to a low-point biserial correlation or DIF are reviewed by our content experts

and excluded from test assembly if the content experts concur that the item may fail to discriminate between ability levels or lead to bias. Furthermore, post-hoc analyses are conducted after the administrations and flagged items are reviewed again by content experts (see Chapter 8 for more details).

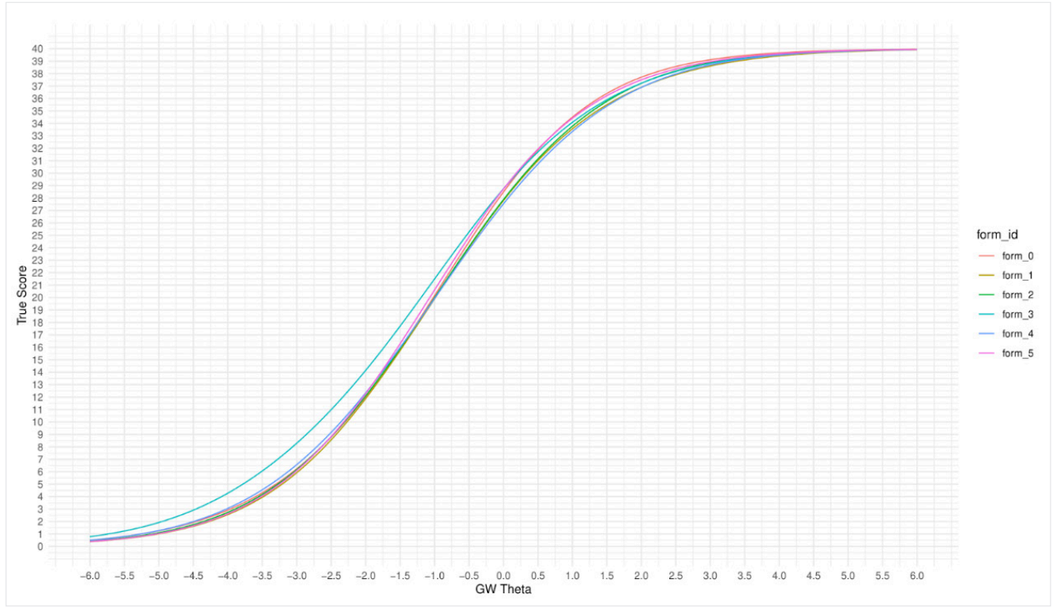
In addition to ensuring that only high-quality items are used, ATA allows us to construct forms that have a consistent level of difficulty. This is accomplished by defining an *objective function*, which is the statistical outcome that the ATA algorithm strives to achieve. For example, test information can be maximized at a given ability level or a certain difficulty level can be targeted. Then, the software finds the combination of items that minimize the differences between the target difficulty and the difficulty of the forms while satisfying the content constraints. The items are pulled from an item bank that is maintained and updated by our Test Development team and psychometricians. Item difficulties are estimated using IRT, which is discussed in Chapter 8. Passage difficulties are estimated based on 1) the difficulties of the set of items associated with the passage and 2) the difficulty of the text of the passage itself.

Figure 3.1 shows the test characteristic curves (TCCs) of the 6 modules assembled for the spring semester of the 2025-2026 academic year. A TCC shows the expected number-correct score on a form given an ability level and the item difficulties. The abilities are on the logit scale as explained in Chapter 8. Each curve in the plots is the TCC of a single module. Figure 3.2 shows the TIFs of the same 6 modules with ability levels on the x-axis and test information on the y-axis. Higher overlap between the curves means that the modules are closer to each other in difficulty and reliability at a given ability level. Given that there are a finite number of items from which the modules can be created, it is challenging to assemble forms that are identical in difficulty. Chapter 8 explains how our scoring process adjusts for the differences between the forms to ensure that scores obtained from different forms are on the same scale and can be compared.

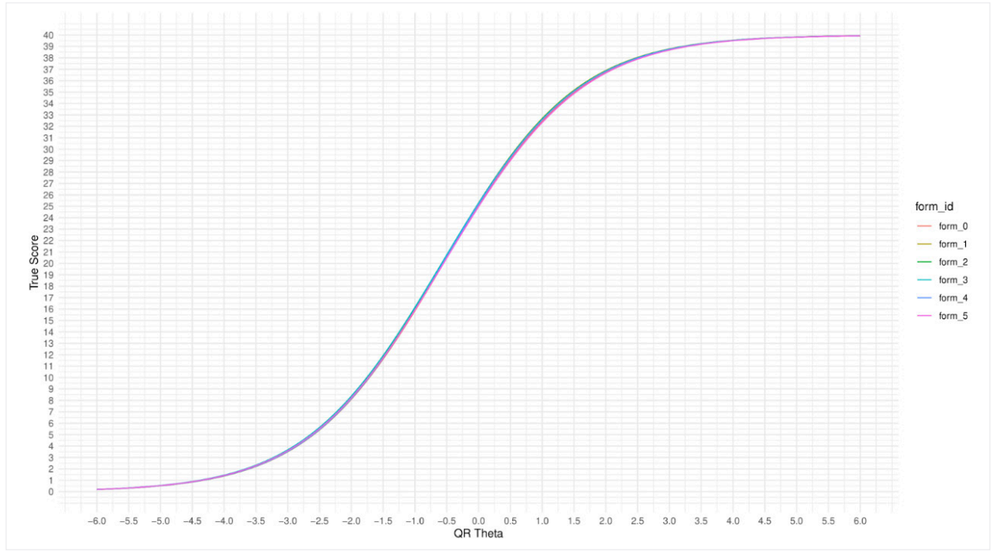
Figure 3.1 *Test Characteristic Curves (TCCs) of the Spring 2026 Forms*



a) The test characteristic curves of the Spring 2026 Verbal Reasoning modules. The x-axis shows a range of ability levels on the logit scale. The y-axis shows the expected score of students with a given ability level. Each line represents the expected scores on a single form.



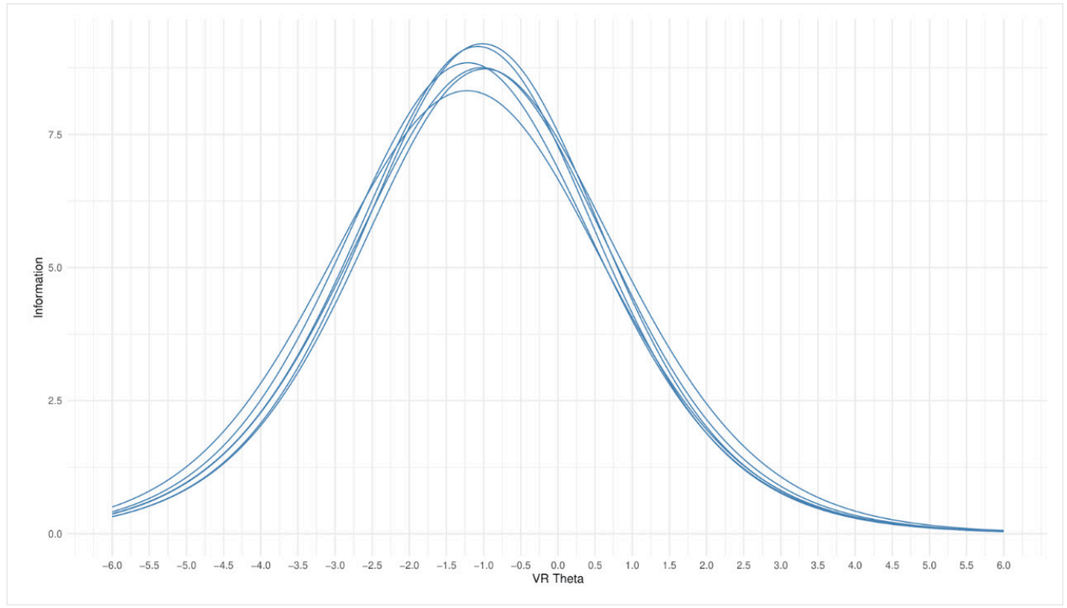
b) The test characteristic curves of the Spring 2026 Grammar/Writing modules. The x-axis shows a rangeof ability levels on the logit scale. The y-axis shows the expected score of students with a given ability level. Each line represents the expected scores on a single form.



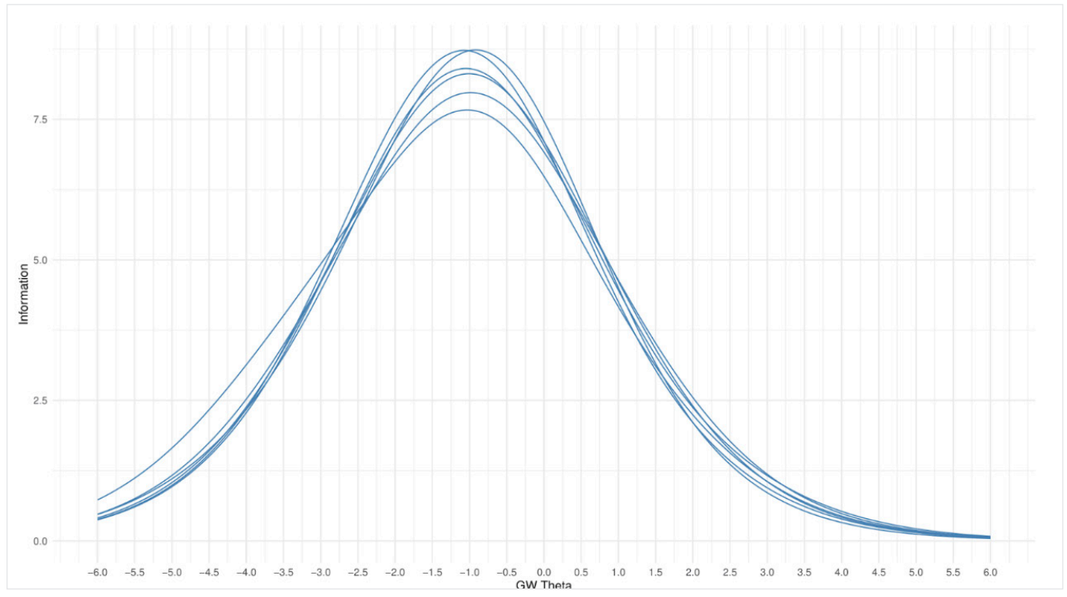
c) The test characteristic curves of the Spring 2026 Quantitative Reasoning modules. The x-axis shows a range of ability levels on the logit scale. The y-axis shows the expected score of students with a given ability level. Each line represents the expected scores on a single form.

Figure 3.2.

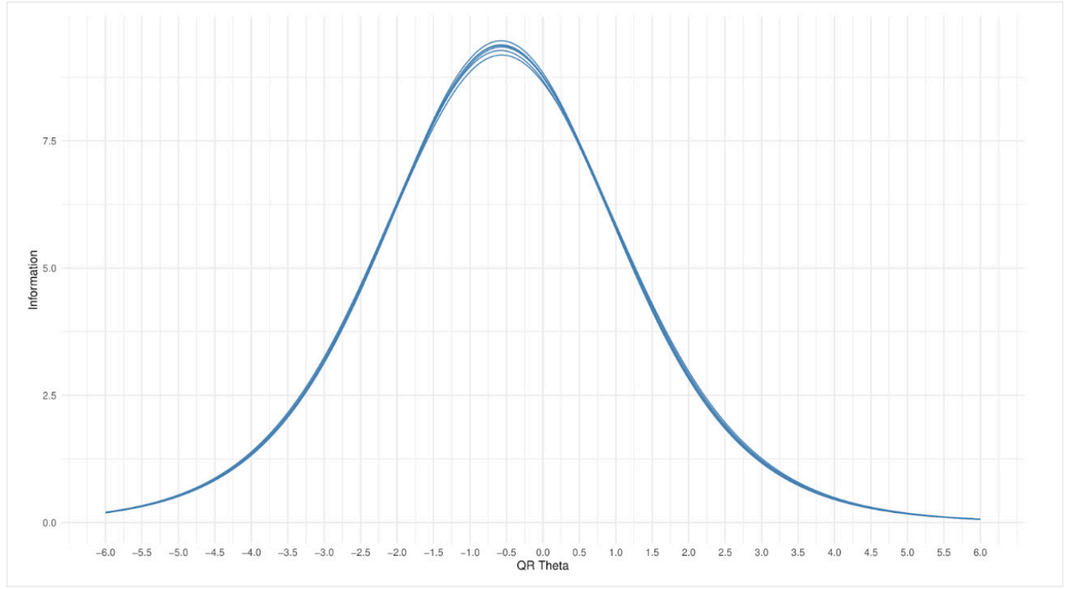
Test Information Curves (TIFs) of the Spring 2026 Forms



a) The test information curves of the Spring 2026 Verbal Reasoning modules. The x-axis shows a range of ability levels on the logit scale. The y-axis shows the total information the items provide about a given ability level.



b) The test information curves of the Spring 2026 Grammar/Writing modules. The x-axis shows a range of ability levels on the logit scale. The y-axis shows the total information the items provide about a given ability level.



c) The test information curves of the Spring 2026 Quantitative Reasoning modules. The x-axis shows a range of ability levels on the logit scale. The y-axis shows the total information the items provide about a given ability level.

Once the parallel test forms have been constructed and the items have been reviewed by the content experts, the passages and the items are uploaded into the test delivery platform. The constituent components of test data in the website user interface are test questions, passages, and images (e.g., graphs, tables, geometrical images). The data is replicated in each field exactly as it is represented in the test Blueprint.

The digital infrastructure for test questions includes variable fields for question numbers (1-120), the text of the question itself, the URL associated with images, the uploaded passage with which the question is associated, the text of answers A, B, C, and D, the correct answer (A, B, C, or D), the difficulty of the question (1-5), and the question type (e.g., "Comprehension—Passage Relationships").

Once all of the passages, images, and test questions are replicated in the website, the online form is reviewed for completeness, correct item ranking, and correct item metadata. Test forms are reviewed to ensure that they meet CLT style according to the House Style Guide. Items are also checked for consistency, typographical errors, correct metadata, and overall coherence of the form. Once the test content has been finalized, the Test Development team completes additional reviews of the test's accuracy and validity. As part of the test development process, proofreaders and editors simulate taking the full test online and in print during each review, which includes checking the answer key, as well as confirming that permissions have been secured for passages.

PAPER TEST FORM

The CLT is primarily an online test, but when a paper version of the test is required, the Test

Development team creates and formats the paper document using the final version of the uploaded test. At this point, the uploaded test may not be changed in any way so that the print form and online form exactly match in regard to content. The paper test is then reviewed in its entirety by a new editor, with a particular focus on formatting, formulas, and other types of errors which might be introduced with the new test mode. With each paper test that is created, a large print version of that test is also created as an available accommodation. As of the 2024-2025 academic year, large print versions are created by request only. The large print test is then reviewed in its entirety by a new editor to ensure that it follows CLT's large print test standards and that no errors have been introduced.

FIELD TESTING

Each section of each form contains up to 10 embedded field test items (25% of the section) pre-calibrated by a machine learning model. The model relies on comparative judgments (Thurstone, 1927) of item difficulty by large language models, together with text-based features for Verbal Reasoning and Grammar/Writing items. The latter include format indicators and lexical frequency measures derived from the *Corpus of Contemporary American English* (COCA; Davies, 2008-), which have been used by Duolingo to predict IRT parameters (Sharpnack et al., 2024). For comparative judgment, LLMs are presented with pairs of items and asked to decide which item is harder. Then, the Bradley-Terry model is used to transform the binary comparisons into continuous difficulty measures (Bradley & Terry, 1952). Comparative judgment has worked well in large-scale human scoring (Office of Qualifications and Examinations Regulation [Ofqual], 2015), and recent work suggests that LLMs also perform substantially better in comparative-judgment settings than in absolute marking, making the same framework useful for streamlining preliminary item calibration (Christodoulou, 2025). The comparative-judgment difficulty estimates are then combined with the textual features to train a supervised regression model on items with known Rasch difficulties. This model is then used to generate preliminary difficulty estimates for embedded field test items before sufficient live response data are available for full operational calibration. After test administration, the final parameters of these items are estimated by anchoring the parameters of the bank items to their known values, and item discrimination and fit are measured using the point biserial correlation and infit/outfit statistics (see Chapter 8 for how flagged items are treated in scoring).

3.4 Quality Control Procedures

CLT employs quality control procedures to ensure that all test items and forms meet established standards for content accuracy, alignment, and clarity. Items are reviewed during development and after administration to evaluate their performance and alignment with intended constructs. This includes analysis of item difficulty, discrimination, and response patterns.

Quality control procedures are designed to ensure that all items and test forms meet established standards for content accuracy, alignment, clarity, and psychometric performance. Items undergo multiple stages of review, including content review, editorial review, and statistical analysis following pilot or operational use.

Decision rules are applied to identify items requiring revision or removal. These may include indicators such as low discrimination, unexpected response patterns, or misalignment with intended constructs. All quality control decisions are documented to support transparency and continuous improvement.

ITEM POOL

CLT maintains a dynamic item pool in order to track individual item use. The Test Development department manages the item pool using a proprietary test content platform that is developed and maintained in collaboration with the Product and Technology teams. The platform features tools conducive to managing the content quality and security considerations of thousands of items in the suite of CLT assessments. This allows the Test Development team to conduct periodic and ad-hoc audits of the item pool using item metadata, as well as to restrict access to items that can be viewed and/or edited by a given item reviewer at any given time. The CLT item pool is secure and only accessible in full to employees with privileged access, ensuring that no active items are made available to students outside of test day. Items with past administration history and known item parameters are also locked for edits, preserving the psychometric integrity of the content over time for future reuse. Any low-quality or problematic items that should no longer appear on a test as part of the active item pool are deactivated and locked but remain in the database for reference and recordkeeping purposes.

3.5 Licensing and Permissions

CLT uses texts drawn from a wide range of literary, historical, philosophical, and scientific sources. These texts are selected to reflect meaningful and enduring ideas and are used in accordance with applicable copyright and licensing requirements.

Permissions are obtained as necessary to ensure that all materials are used appropriately and in compliance with legal and ethical standards. CLT prioritizes the selection of passages that are in the public domain in the United States; however, licensing is sometimes required for passages that are not yet in the public domain. In these cases, CLT contacts the rights-holders directly to negotiate terms of use.

3.6 Content Review and Editorial Review

All items undergo a structured content and editorial review process prior to inclusion in operational test forms. Content review focuses on alignment to the test blueprint, accuracy of subject matter, and appropriateness of source material. Editorial review ensures clarity, consistency, and adherence to established style guidelines.

Reviewers evaluate items for clarity of language, appropriateness of difficulty, and alignment to intended constructs. Items that do not meet established criteria are revised or removed from consideration. This multi-stage review process supports the development of high-quality items that are both rigorous and accessible to the intended population.

TEST FORM DEVELOPMENT

Test form development proceeds as described in Section 3.3. Quality control procedures for each test form consist of checks at every stage for consistency with the blueprint and overall style of the test. At form construction, validation that the blueprint has been met is provided as an output of the automated test assembly algorithm and is confirmed via a separate check performed by the Test Development team. Any revisions to the form are reviewed and approved before the form becomes operational. The various

stages of content development and review are summarized below.

PASSAGE SELECTION (VR/GW SECTIONS ONLY)

Reading passages for the CLT are sourced and selected by the Test Development team. Passages must meet word count and category specifications (see Section 2.3 and Section 2.4), and must be coherent excerpts that are representative of the source work and author but not so ubiquitous as to be broadly familiar to test takers. In addition, passages are evaluated for appropriate difficulty using a combination of human judgement and algorithm-based analysis of text complexity.

Passages that meet the above requirements and/or are sourced from the Author Bank are not considered automatically approved by Test Development leadership; passages are ultimately selected and approved based on the merits of their content. Passages are thoroughly evaluated for worthiness of inclusion on the CLT on the basis of truth, goodness, and beauty, in accordance with CLT's mission to provide meaningful assessments (see Section 1.3).

ITEM DEVELOPMENT

Writing of test items may be performed by internal Test Development staff or by contracted content experts. All Test Development contractors are provided with item specifications according to the test blueprint, content standards by subdomain and item type, and CLT's style guide. In addition to resources tailored to each specific assignment, all item writers have access to a library of reference materials curated by the Test Development team to support item quality at every stage of writing and review. Item writers create their item content directly within CLT's test content platform, which supports a variety of rich text formatting options, MathML and Ascii Math input modes, and images in both question stems and answer choices as needed. CLT's content management platform also includes a dedicated field for item developers to include an item rationale and distractor analysis commentary on each item.

DEVELOPMENTAL EDITS

After new items are written, they undergo an initial round of review to be edited and reworked, ensuring the item is in alignment with CLT's content and style standards. Items may also be adjusted at this stage with the intent of increasing or decreasing the expected difficulty, eliminating any anticipated parameter concerns such as low point biserial, and proactively resolving any other potential answer key issues (e.g. multiple or no key).

INTERNAL FORM SCAN

This is the first round of form-level review following automated test assembly, which typically occurs after new content has already been developed so that new content may be included. An internal member of the CLT Test Development team reviews the form in full, performing validation checks to ensure that the form represents a cohesive test that does not deviate from the blueprint, precedent, or general best practices. For example, the passages appearing together on the form are evaluated as a whole to confirm that there is enough variety of topics addressed and/or author time, place, and viewpoint included so as to avoid passages that could clue each other's items or interact in other ways that could affect response data. The internal reviewer may also need to adjust item ranks at this stage in order to avoid too many items of the same Question Type or Subdomain occurring consecutively, maintain any snob item relationships,

ensure passage-based items occur in a reasonable order, etc.

Any new content on the form (up to 25% of each section) is also thoroughly evaluated again by internal Test Development staff to ensure the new items do not need additional developmental edits before they can be presented on a test and/or be calibrated in the IRT model. Special attention may also be given to items that have not been administered in 3+ years in order to maintain CLT's latest standards for content or style. The internal reviewers also perform final checks for any enemy items on the form at this stage and work with the Psychometrics team to replace any items that should not appear on the same form together.

CONTENT EDITS/PROOFREAD

The second form-level round of review is completed typically by a contracted item reviewer. This round of review primarily entails the reviewer(s) simulating taking the full test online and checking the answer key Reviewers must confirm that the keyed answer is correct and that all distractors are incorrect.

FINALIZATION

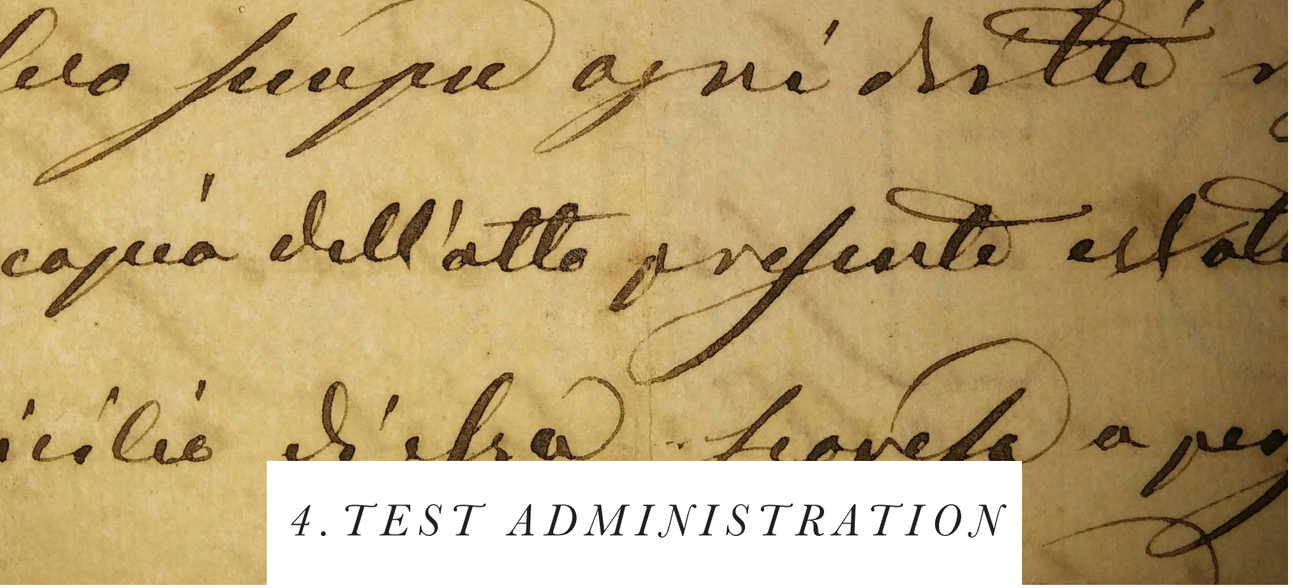
Finalization procedures include an additional answer key check, an accuracy check of all supplementary question information and item metadata, and a check that all intended edits noted in earlier rounds of review were made in the database. Passage licensing information is also verified to confirm that the rights for any passage not in the public domain have been appropriately secured and that all passages are credited correctly.

FORM SIGN-OFF

Once a form has completed all prior rounds of review, a member of Test Development management accesses the form through the test platform to view the test as it will appear during live administration and authorize any final changes. The designated manager then completes a sign-off process in the test content platform which locks the form and its items from any further edits, and the manager verifies that they are approving the form for administration. After Test Development form sign-off is complete, the Operations completes an additional platform review of all form content before administration.

FLAGGED ITEM REVIEW

During the scoring process for each test administration, the Psychometrics team compiles a list of items with response data or parameters that did not meet expected thresholds. These may include poor fit to the IRT model (e.g. low point biserial, high infit or outfit), difficulty that is too high or too low to effectively measure the test population (extreme p-value), response data that contradicts the item's previously-calibrated difficulty (beta drift), etc. The Test Development team reviews the psychometric data and content of these items in tandem, and makes determinations on how to proceed with the flagged items on a case-by-case basis. Items may be retained as active items in the pool with no changes; they may require the creation of a related but new item with revisions to the question stem, answer choices, or both; or they may be retired altogether. This item flagging process ensures that all active items within the item bank adhere to CLT's item quality standards. More details on item analysis and flagging can be found in Chapter 8 of this report.



4. TEST ADMINISTRATION

4.1 Overview

The CLT is offered multiple times per year. The test is normally administered to students online, either at a user-selected private location (typically at home, but sometimes in a private room inside a public facility, such as a library) or at a CLT partner school (for schools that contract with CLT to administer the exam "in-house" to their students). Schools who administer the CLT have the choice of administering the test either online or on paper.

The test is proctored remotely when administered privately; CLT staff record and review the tests to ensure exam integrity. In-school CLT administrations are proctored by school staff.

Students receive two hours to complete the CLT: 40 minutes for the Verbal Reasoning section, 35 minutes for the Grammar & Writing section, and 45 minutes for the Quantitative Reasoning section.

If the exam is taking place at a CLT partner school, the proctor ensures that students proceed from section to section together. In private administrations of the test, students may move on early if they choose (up to and including submitting the exam early), but still must move on once the timer for that section expires. Students cannot return to a previous section at any point, in either form of administration, and time "saved" on one section cannot be transferred to another.

Any difficulties that arise during an in-school exam will normally be handled by the proctor. For students who run into problems while testing privately, our live chat support is available to assist them throughout the day; there is no test-time penalty for consulting chat support.

Testing accommodations are available for students with documented disabilities. These may include extended time, extra breaks, use of a calculator, or other policy modifications, as necessitated by the student's disability. Accommodations are described further in Chapter 5 of this report.

4.2 Test Modes

The CLT is administered in two different online modes and one paper mode to make testing convenient and secure for all students.

IN-SCHOOL MODES (ONLINE & PAPER)

In-school testers may take the CLT as an online exam or on paper with an answer sheet. The school will register for the test and order their tests prior to test day. Students may not register directly for an in-school test or take this online version of the exam from their home.

For in-school tests, the school administering the test provides a trained proctor for the exam for both the online and paper modes of the test. This proctor will provide specific test day directions and guide students through the test. The proctor is also responsible for contacting CLT in the event of a technical issue on test day.

ONLINE

Students taking the CLT from a school as an online exam will use a laptop, desktop computer, or tablet. Students will normally bring a laptop or tablet for their own use. Some schools may choose to provide suitable devices for all students taking the test.

The online test will work on most modern devices. It requires a reliable internet connection with Javascript enabled, and students must have LockDown Browser® downloaded on their devices. Questions in the Quantitative Reasoning portion of the exam may include mathematical notation. Mathematical notation is rendered with MathJax.

Once the test is complete, proctors and administrators will complete a post-test survey about their test experience and note any anomalies during the exam administration. Scores for online exams are released the within one week of the administration.

PAPER

Students taking the CLT on paper will receive a test booklet and answer sheet. They will fill their answers out in the answer booklet, along with their identification information. The optional essay is not available on the paper test.

School administrators order their exam kits at least 6 weeks prior to the test administration date. The kits, which include exam booklets and answer sheets as well as instructions, are mailed to the school a minimum of one week ahead of the test date. They are sent to the attention of the school's primary point of contact. As with the online CLT, proctors are expected to follow a strict process, outlined in the paper test manual.

Once the test is complete, schools return the answer sheets to CLT for processing. Students and administrators receive their scores and analytics within 30 days of the return of the answer sheets.

AT HOME MODES (REMOTELY PROCTORED TEST)

The remotely proctored test is a convenient choice for homeschooled students and students whose schools do not yet offer the CLT. For students that attend CLT partner schools, it is also a good way to get ready for an in-school test administration.

To test at home, students must create a profile on the CLT website and sign up for the specified exam date. Once registration and payment are completed, the student receives instructions on how to prepare for the remotely proctored test, including setting up their space, checking their internet speed and computer settings, and simulating a test. On test day, students sign into their profile to access the test.

Students interested in taking the remotely proctored CLT at home do so using their own desktops or laptop. If necessary, students can take the test from another location such as a library, church, or a friend or relative's home.

For the remotely proctored test, no onsite proctors are required: test integrity is maintained through CLT's test administration software. The student must test alone in a closed, well-lit room, from the beginning of the exam until it is submitted. The test may be taken any time during the day that the CLT is offered. Live chat support is available during the exam from 7am to 7pm Eastern Time, for students who encounter any difficulties.

Students are encouraged to become familiar with the test requirement and layout prior to testing to ensure a smooth testing experience. A stable internet connection is required, and students must use a laptop or desktop computer with a functioning camera and microphone. Tablets and mobile devices are not compatible with the remote proctoring software. CLT has developed a number of tools to assist students, including troubleshooting guides, instructional videos, and a fully-featured test simulation that operates the same way as the operational exam to allow students to test their system.

CLT requires a photo ID to verify student identity on the remotely proctored test. Additionally, there is no optional essay available on the remotely proctored test. During the exam, the CLT records both the student's screen and their camera to ensure that test integrity is maintained. The exam recordings are reviewed by CLT staff following the test. Testers who are found to have violated the CLT honor statement will not receive scores on the test. Scores for the Remotely Proctored CLT are released the third Wednesday following the administration.

PRACTICE TESTS

CLT provides three full CLTs on every student account. Using these tests, students can become familiar with the format and content of the online test as well as the testing interface. Three additional practice tests are available in hard copy in the Official CLT Student Guide.

4.3 Test Day Processes and Procedures

Students may take the CLT only under secure, supervised conditions.There are two ways that students can take the CLT: on paper or online during an in-school test day at a CLT partner school, or at home with the Remotely Proctored exam.

IN-SCHOOL TESTS

Admitting Students into the Testing Room - On test day, proctors have the final list of CLT students for their specific test site on their CLT accounts. The manual instructs proctors to verify students' identity before admitting them into the testing room, using any of the following types of approved photo ID:

- Passport

- Driver's license or permit (if photo included)
- State ID
- Military ID
- High school ID (current year only)
- HSLDA student ID (current year only)
- CLT Student ID Form

Proctors then assign seats for every admitted student.

Test Access Code - In order to take the exam on test day, students must enter the access code specific to the exam in question. Proctors receive the access code directly from CLT the week before and the day before test day. They provide their students with this access code once all authorized students have been admitted and seated and the preliminary instructions have been read.

Calculators - Calculators are not allowed on the CLT, including on the Quantitative Reasoning section, unless a student has been specifically approved for a calculator as a testing accommodation. Questions are designed to be solvable without the use or need of a calculator.

Timing - Our test delivery platform is equipped with a built in timer to support the students ability manage their progress through each section. Proctors provide students with a Section Access Code in order to progress to the next section. To aid the proctor in determining at a glance whether all the students are working on the appropriate section of the exam, each section is color-coded for the online test. A similar aid is available to proctors of paper exams: the names of the first, second, and third sections are printed in bold at the top-left, center, and right of the pages, respectively.

Anomalies - Proctors must submit the Test Day Anomaly Report to CLT before exiting the testing room. They are instructed to note any testing anomalies on this report. Instructions for potential testing anomalies that are to be noted on the report include:

- Students who do not arrive to an exam
- Students who arrive late to an exam
- Students who leave during an exam
- Students who use an additional device
- Students who become ill during an exam
- Questions asked during an exam
- Disturbances during an exam
- Emergency evacuations
- Power failure
- Wifi failure
- Device failure
- Site failure
- Copying test materials

PROCTORS

Proctors are responsible for ensuring that the in-school exam is administered and taken under the highest security standards possible. CLT proctors must be at least 21 years of age and cannot be related to the students they are proctoring. Each proctor monitors no more than 20 students, allowing for differences in room size and layout. During the exam, the proctor must be able to see all students and ensure that the spacing requirements are respected. Proctors may not provide assistance to students on exam content.

It is the proctor's responsibility to administer the exam fairly, safely, and securely. In order to do so, proctors are responsible for the following duties:

1. ***Setting up for the Exam:*** Prior to the exam, proctors prepare the room for testing according to the guidelines. Proctors also assist students with filling out their identifying information on their test sheets as needed.
2. ***Monitoring Students:*** Proctors ensure that no students access any of the following prohibited items:
   - » Cell phone or other device (must be completely off and out of sight)
   - » Calculator
   - » Digital watch with internet access, communication capabilities, or calculator
   - » Books
   - » Resource/reference material of any kind
   - » Snacks (may only be eaten during the ten-minute break).
3. ***Remaining in Testing Room:*** With the exception of the restroom break and emergencies, students must remain in the testing room for the duration of the test. Proctors are not allowed to leave students alone during the exam, even before the exam has begun.
4. ***Maintaining Exam Security:*** All CLT exams are copyrighted and cannot be copied, printed, or otherwise used outside of the test. Proctors may not alter CLT materials, transfer them to another file, or make copies. They also may not disclose test materials, questions, or other information to any outside parties. Proctors are tasked with protecting the content of the exam by ensuring that students do not copy or otherwise duplicate exam material, such as by taking pictures of their tests.
5. ***Completing Test Day Anomaly Report and Proctor Survey:*** Immediately after the exam, proctors should fill out and submit an Test Day Anomaly Report and Proctor Survey which notes any anomalies that may have occurred. CLT staff review these reports as well as testing data and may follow up with school administration as needed.

REMOTELY PROCTORED TESTS

The Remotely Proctored CLT is administered privately and without a proctor; CLT staff record video, screen, audio, and keystrokes during the test, and review it afterwards to ensure exam integrity. Recordings are stored in a secure location and deleted within 30 days.

Access to the exam is emailed to the test-taker and their emergency contact the evening before test day. On test day, a student logs into their account when ready, and once their profile is complete, they start the test from the student dashboard: they enter the Test Access Code, read and sign the Honor Code, and

complete their pre-test instructions. The timer does not start until the first section of the test is begun.

Technical and customer support is available from 7am to 7pm Eastern time on test day. Students are strongly encouraged to test during these hours. The test must be taken in one sitting. The test is open from 12:00 am to 11:59 pm Pacific Time on test day. The exam takes about two hours and twenty minutes, including pre-test instructions and procedures. Students will not incur any time penalties for chatting with CLT support during the exam.

TESTING ROOM REQUIREMENTS -

1. Students must be alone in a closed, well-lit room from the beginning of the recording until the test is submitted. Public spaces such as libraries, cafes, or parks are not allowed. If it is not possible to meet this requirement, students must contact CLT with details and we will do our best to arrive at an acceptable arrangement.
2. Students must remain in the room alone with no talking throughout the test. Students are asked to post the CLT Remote-Proctored Test Sign as a reminder to other members of the household not to interrupt. Before starting the test, students should call or text anyone who might come home while they are in the midst of testing.
3. Students should be in a room with a reliable internet connection, preferably as close as possible to the Wi-Fi router.
4. Students' computers and keyboards must be on a desk or table.
5. Students must sit on a standard chair or stool (not a bed, couch, or overstuffed chair).

REQUIRED ITEMS

1. A laptop or desktop computer with a functioning camera and microphone.
   - » Tablets and mobile devices cannot be used.
   - » Both internal (built-in) and external (e.g. USB) cameras and microphones are acceptable.
   - » Students must make sure their computer's speakers are working and turned on so that they can hear the notification tones for the test timer.
   - » If using a laptop, students must make sure it is plugged in during the exam.
   - » LockDown Browser® must be installed on the student's computer.
2. An approved form of photo ID.
   - » Passport, driver's license or permit, or state ID
   - » High school ID (current year only), HSLDA Student ID (current year only), or college ID
   - » Military/military dependent ID
   - » If students do not have any of the above, they may print the CLT Student ID Form and have it notarized by a notary public, or signed and sealed by a school official.

4.4 Test Day Schedules

The CLT must be completed in the order and time given. In-school testers taking the CLT must remain for the full time of each section and submit their exams simultaneously with the other students present,

even if they finish one or more sections early.

Testers taking the remotely proctored CLT may move to the next section early (including submitting the exam early) if they finish with extra time. The remotely proctored test contains a test timer and once time has elapsed for a section, students are no longer allowed to enter or change answers.

For in-school tests, proctors are responsible for each of the test sections and providing instructions for test takers. The entire test administration will take the proctor about three hours if no students take the essay, or about three hours and thirty minutes if at least one student takes the essay.

SAMPLE SCHEDULE (IN-SCHOOL TEST)

TIME	TASK
9:40 AM	Proctor gathers required items and prepares the testing room.
10:00 AM	Proctor admits students and reads General Announcements.
10:10 AM	Proctor reads Administrative Material.
10:20 AM	Section 1: Verbal Reasoning begins.
11:00 AM	End of Verbal Reasoning section, beginning of Grammar/Writing section.
11:35 AM	End of Grammar/Writing section, beginning of restroom break.
11:45 AM	End of restroom break, beginning of Quantitative Reasoning section.
12:30 PM	End of Quantitative Reasoning section; closing announcements and student surveys.
12:35 PM	Dismissal of students not taking optional essay, beginning of essay for remaining students.
1:05 PM	End of the optional essay, dismissal of remaining students.
1:10 PM	Proctor submits Test Day Anomaly Report and Proctor Survey.

4.5 Test Day CLT Support

Live test-day support for proctors, administrators, and testers is available on test day. CLT has a dedicated team of customer service representatives who are available to answer questions from schools, proctors, and parents.

This team includes representatives from CLT's technology, operations, and customer support teams to ensure that issues can be resolved quickly and directly. On test day, live support is available via live chat and phone call.

4.6 Test Security

Classic Learning Initiatives (CLI) test security is designed to ensure the privacy of its test-takers.

The management of their data is described below.

DATA SECURITY

CLI trains all its employees on the high sensitivity levels of CLT data, including the access and use of confidential material such as personally identifiable information (PII). CLT requires each employee to acknowledge and sign internal policies regarding the acceptable use of CLT data. Our security measures are annually reviewed by a third party to ensure we are meeting external standards of data protection.

DATA PRIVACY AND ACCEPTABLE USE

CLT considers all student data confidential, including collected identifiable information (email and student profile data) as well as test results. CLT employees may not share any student's data with a third party without that student's express consent.

Students who take their tests through their school will have access to their scores and analytics. Their scores and analytics will also be available to school administrators, and teachers. All students may opt to share their profile and test results with specific colleges of their interest and/or opt into CLT's partnership program in which CLT shares limited student data with partner institutions. Students who opt in may also opt out of the program at any time by logging into the CLT web application and editing their profile.

Proctors can view limited student data on test day to facilitate the test and verify attendance. Proctors do not have access to a student's full profile, test history, or any other data. Proctors are not permitted to share any student information with any third parties.

School administrators can view full student data for test day, including test history, scores, and basic profile information. School administrators do not have access to the full student-entered user profile and cannot view student score shares, practice tests results, or independent registrations or purchases.

ACCESS CONTROL

CLT data may be accessed either through the web application or through the database directly. All users must be authenticated to access CLT data, and authorization is based on security level.

- Web Application Access – The CLT web application security is role-based. By default, all users who register for an account receive the same level of access as students, the most minimal access level.
  - *Support Access* – CLT employees are granted a support role in order to access necessary information to support customers. Users in a support role can view test registrations and view student data, but they cannot access the test management section of the application.
  - *Privileged Access* – a limited number of CLT employees have privileged access that allows them access to write, review, and modify test data in advance of test dates. This includes the ability to add tests, add and edit questions and answers in existing tests, change test dates and deadlines, and deactivate tests. Privileged access users are required to sign an additional policy regarding test integrity and the acceptable use of test data. Privileged access may be granted only by the Chief Technology Officer.
- Database/Network Access – accessing the database directly falls under privileged access and is limited to the development and analytics teams. Network traffic to access the database is restricted by IP address. Each privileged user is granted two accounts, one read-only and one administrative account. Users use their read-only account unless a critical change is required. Some users, such as

those on the Analytics team, may be granted only a read-only account.

» Data Access – all CLT data is stored in a secure cloud environment that is not accessible to CLI employees in general, only to authorized members of the technical and operation teams. The third-party cloud provider ensures the highest level of security and access.

MONITORING AND AUDITING

All activities are logged when changes are made in the software, database, and infrastructure. Logging is monitored on a regular basis to identify breaches, risks, or unexpected behavior. User roles are also monitored on a regular basis to ensure that users have not been inappropriately granted access to data.

INCIDENT MANAGEMENT AND RESPONSE

The CLT Executive Team manages all incidents, including data breaches and/or unacceptable use of data. In the event that user data is compromised, the issue will be immediately remediated and the affected parties will be contacted. CLT also conducts an after action report that is submitted to a third party for evaluation.

5. TEST ACCESSIBILITY

5.1 Fairness During the Testing Process

All CLT testing takes learning differences and disabilities into account, in accord with the Standards for Educational and Psychological Testing (*Standards*) jointly set forth by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. CLT also considers fairness in testing a top concern, and persistently works to minimize bias and ensure a universally accessible design.

Using language from the *Standards*, we begin to define fairness as accessibility: "the notion that *all* test takers should have an *unobstructed opportunity* to demonstrate their standing on the construct(s) being measured."[1]

Testing accommodations are adaptations to an exam that can be made for students with diagnosed disabilities; their purpose is to provide candidates with full and equal access in order to accurately demonstrate their skills and abilities as measured on the test. (Accommodations on the CLT do not guarantee test completion, improved performance, or any other specific outcome.) All testing accommodations are made on a case-by-case basis. Regardless of diagnosis, we ask that individuals seeking disability-related accommodations provide us with documentation of the nature of their disability and its relevance to the test. Accommodations for the CLT must be submitted for approval at least four weeks prior to the test administration date.

1 American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). Standards for educational and psychological testing. American Educational Research Association.

5.2 Fairness in Test Accessibility

CLT provides testing accommodations to students with documented disabilities to make testing equally accessible to all. Test accommodations are individualized and considered on a case-by-case basis.

Regardless of diagnosis, all individuals seeking disability-related accommodations must provide evidence that their condition rises to the level of a disability, which adversely affects a child's educational performance, and provide information about those functional limitations. Demonstrating that an individual meets diagnostic criteria for a particular disorder does not automatically mean that the person qualifies for test accommodations. Accommodations must be appropriate to the particular task and setting involved, and proper documentation of the effective use of accommodations in classroom or individual learning activities should support use in testing.

5.3 Accommodations and Requests

CLT is committed to providing every student a fair test-taking experience by ensuring the security, integrity, and validity of its examinations. CLT is committed to providing access to its programs and services to students with documented disabilities, a disability being a physical or mental impairment that substantially limits a major life activity.

CLT therefore offers a range of accommodations for students with documented learning or physical disabilities, in accordance with the Individuals with Disabilities Education Act (IDEA) and the Americans with Disabilities Act (ADA). In compliance with these laws ,and in keeping with its efforts to provide equality of access to the test, the CLT seeks to minimize bias and promote cognitive diversity. Beyond these laws, we also offer ELL accommodations.

Test-takers seeking accommodations are required to submit an accommodations request. Information is available on the CLT website. Accommodations approvals are granted for a time period of up to five years.

All accommodations requests must be submitted on behalf of individual students at least four weeks in advance of the testing date. An Accommodations Request Form submitted for more than one student will not be considered.

When accommodations requests are submitted by school administrators on behalf of individual students, parents must also submit a Consent Form for Releasing Accommodations Documentation which authorizes the student's school to release accommodations-related documentation to CLT.

Approved accommodations on the exam may include:

EXTENDED TIME

- 25% Extended Time
- 50% Extended Time
- 100% Extended Time

MEDICAL NEEDS ACCOMMODATIONS

- Food/drinks/medication in the test space
- Medical devices in the test space
- Further monitoring, if requested
- Ability to pause the timer, if needed, to adjust blood sugar levels

ELL STUDENT ACCOMMODATIONS

- 50% Extended Time
- Approved bilingual word-to-word glossary

CALCULATOR

- 4-Function Calculator. No scientific or graphing calculators are permitted.

MISCELLANEOUS

- Text to speech
- Reader
- Scribe
- Read aloud to self
- Breaks between sections
- Additional scrap paper
- Large font exam
- Small group testing
- Other accommodations can be approved and provided as needed for access to the exam.

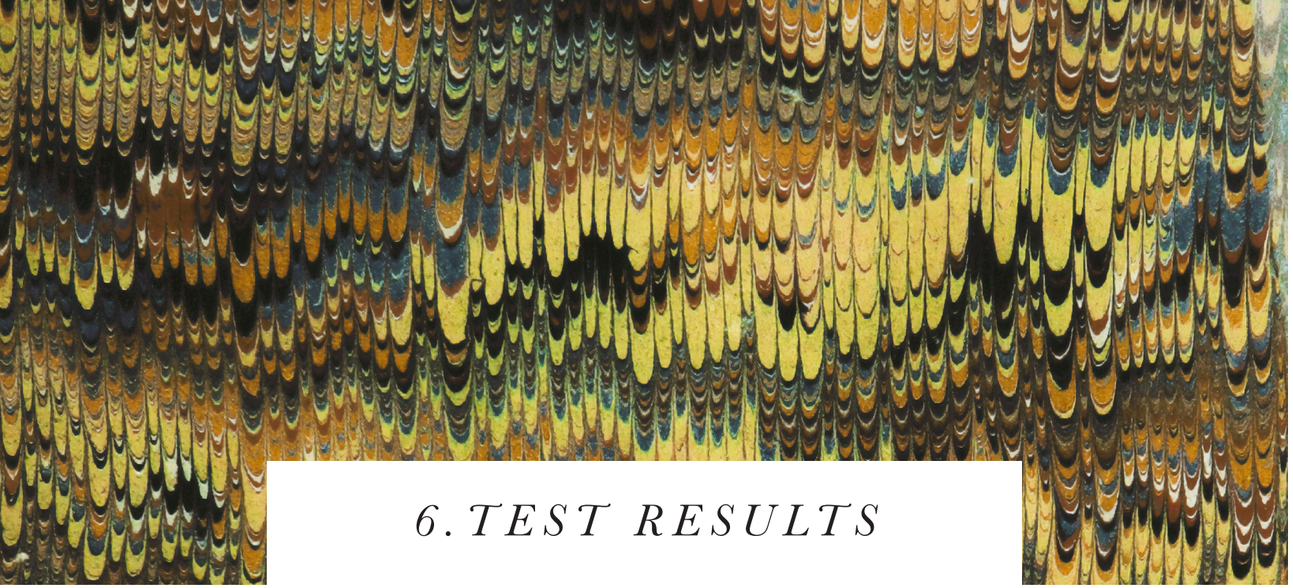
REVIEW TIMELINE

To ensure the timely fulfillment of accommodations requests, such requests must be submitted (with supporting documentation) at least four weeks before the test date.

CLT reviews accommodations requests and submitted documentation and will contact the submitter about any matters requiring clarification. Please note that if a request is incomplete when uploaded, it may take longer to process while we request the required documentation. CLT keeps the submitter updated as to the status of their request.

CLT staff will make every effort to review and approve requests; however, CLT cannot guarantee a full review for requests received after the accommodations deadline. In order to be fair to all candidates, accommodations requests are reviewed in the order they are received; requests cannot be expedited.

Testers may appeal an accommodation decision if their request is not approved. Successful appeals should include a specific reason for appeal, as well as additional documentation beyond what was included in the original request.



6. TEST RESULTS

6.1 Student Score Reports

Students receive test results as part of a score report which is available to them through their online accounts on the CLT website. The data provided helps students and teachers identify the areas on which a tester should focus. CLT score reports may also be shared with partner colleges as part of the college admissions process.

An individual student score report has five main sections, as pictured and described below.

1. SCORE SUMMARY

This part of the Student Score Report shows the CLT scaled score on the overall test and on each section. The overall scale ranges from 0-120 and the sections contain scaled scores from 0 to 40. Testers also see a concorded score on the SAT and ACT as well as a national percentile, which allows a tester to compare their projected score to the scores of a nationally representative group on the same test.

Score Summary

Scores are shown by subject area and total.

	Adjusted Score
Overall Score	97
Verbal Reasoning	38
Grammar / Writing	39
Quantitative Reasoning	20

SAT / ACT Concordance	Projected Score	Nat'l Percentile
SAT	1420	98th
ACT	32	N/A

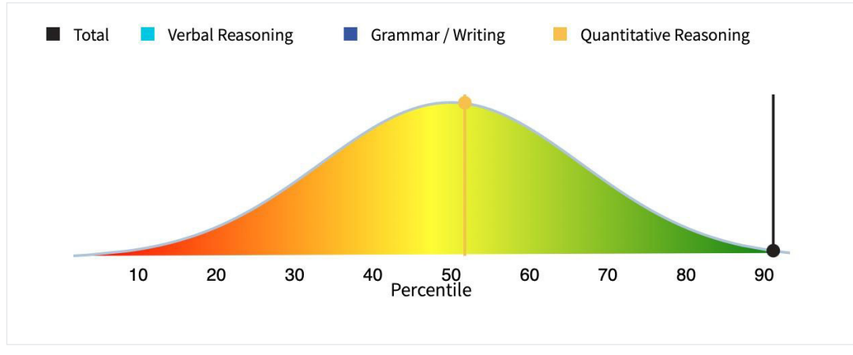
2. CLT USER PERCENTILES

CLT User Percentiles show the percentage of CLT scores that are equal to or below the tester's score.

	CLT User Percentiles ⓘ
Overall Performance	91st
Verbal Reasoning	99th
Grammar / Writing	99th+
Quantitative Reasoning	51st

3. CLT SCORE BELL CURVE

The bell-shaped figure visualizes the distribution of all CLT scores. The black line locates the user percentile of the tester's total score on the exam. Scores in the yellow zone are average, while scores in the green zone are above average.



4. DOMAINS AND SUBDOMAINS

The Domains and Subdomains Report shows tester strengths and potential weaknesses. "Top Question Types" shows the types of questions which were answered with the highest accuracy.

Top Question Types

Question Type by Academic Domain-Subdomain Pairing	% Correct
Grammar - Agreement	100%
Writing - Word Choice	100%
Grammar - Punctuation and Sentence Structure	100%
Writing - Style	100%

In contrast, "Areas for Improvement" shows the types of questions with the lowest percentage of correct answers.

Improvement Areas

Question Type by Academic Domain-Subdomain Pairing	% Correct
Geometry - Trigonometry	0%
Geometry - Plane Geometry	25%
Geometry - Properties of Shapes	50%
Mathematical Reasoning - Logic	50%

From this section, testers may also access more information about each question's subdomain and view example problems in that category.

5. DETAILED PERFORMANCE

Below the Top Question Types and Areas for Improvement, testers may access a detailed view of your performance on each subject, domain, and subdomain. The "% Correct" column shows the percentage of questions you got correct in each category.

Detailed Performance by Subject Section, Academic Domain, and Academic Subdomain

Question Type by Academic Domain-Subdomain Pairing	% Correct	Example Practice Questions
− **Verbal Reasoning**		
Analysis	**100%**	
Interpretation of Evidence	100%	12
Textual Analysis	100%	4
Comprehension	**93%**	
Passage as a Whole	100%	8
Passage Details	82%	6
Passage Relationships	100%	10
+ **Grammar / Writing**		
+ **Quantitative Reasoning**		

6.2 College Score Reports

The student has the option to share their CLT score report with as many colleges as he or she chooses at no additional cost. If the student completes the optional essay section, he or she may also choose whether or not to share the text of the essay with colleges.

When students opt to send their score reports and optional essays to colleges of their choice, these colleges receive those students' CLT score information, as well as their projected ACT and SAT score based on our concordance chat.

CLT Official Score Report

Student Information

Name	Example Name
DOB	2006-01-01
Address	Example Address Annapolis, MD 21401
Email	example
Phone	example
Gender	example

CLT Scores

Test Date	October 15, 2022
Verbal Reasoning	38 / 40
Grammar/Writing	39 / 40
Quantitative Reasoning	20 / 40
Total Score	97 / 120

School Information

GPA	3.9
High/Home School	
High/Home School Type	Homeschool
Graduation Class	2024

Additional Information

Projected SAT	1420*
Projected ACT	32*
Intended Major	Arts and Humanities
Financial Aid Interest	Not Provided

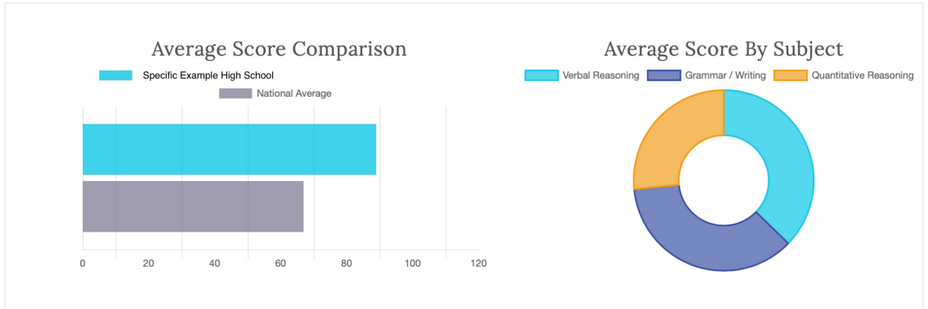
* Projected based on concordance with CLT score

6.3 Secondary School Score Reports

CLT provides detailed class and individual analytics for schools who have offered the CLT exam.

Once scores for an exam have been released, administrators of secondary schools and home school organizations may view student scores by logging in to their CLT school administrator accounts to view scores and analytics.

Students can view their own scores by logging in to their CLT accounts and viewing their individual student score reports, as described above.



Analytics include historical average scores for the school, as well as scores and CLT percentiles for each student, per test. CLT percentiles are user-referenced and indicate how a student performed on the test as compared to their user group.

Students (Test Date)				Overall		Verbal Reasoning		Grammar/Writing		Quantitative Reasoning		
Last Name	First Name	Grad Year	DOB	Score	CLT User Percentile	Score	CLT User Percentile	Score	CLT User Percentile	Score	CLT User Percentile	Actions
Student 1	clt	2024	2006-01-09	85	70th	30	64th	24	29th	31	94th	
Student 10	clt	2021	2006-05-04	95	88th	30	64th	32	76th	33	97th	
Student 2	clt	2024	2005-10-23	93	85th	32	76th	32	76th	29	90th	
Student 3	clt	2023	2004-11-30	78	55th	27	46th	26	39th	25	77th	
Student 4	clt	2024	2005-12-22	67	31st	24	31st	21	18th	22	62nd	
Student 5	clt	2024	2005-09-29	79	57th	28	52nd	25	34th	26	81st	
Student 6	clt	2024	2006-03-21	74	45th	23	27th	24	29th	27	84th	
Student 7	clt	2024	2005-10-10	60	19th	20	16th	18	9th	22	62nd	
Student 8	clt	2022	2004-04-01	92	84th	32	76th	33	82nd	27	84th	
Student 9	clt	2021	2003-06-28	68	33rd	21	19th	27	45th	21	57th	

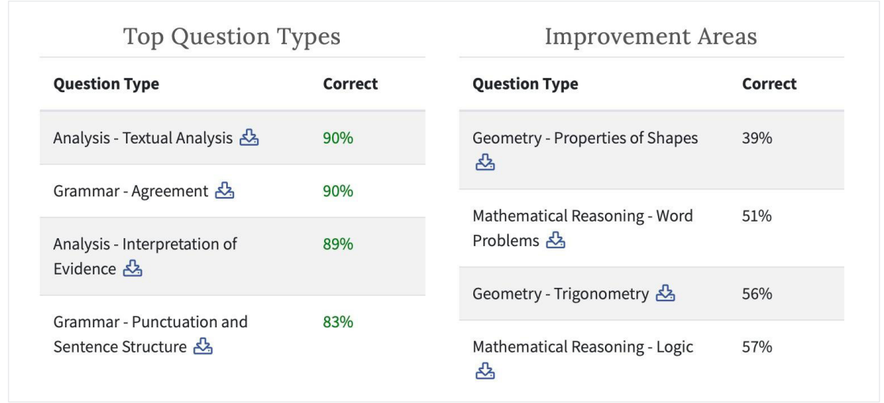
Test administrators also have access to detailed student and school-level analytics. Student performance is reported individually by academic domain and academic subdomain.

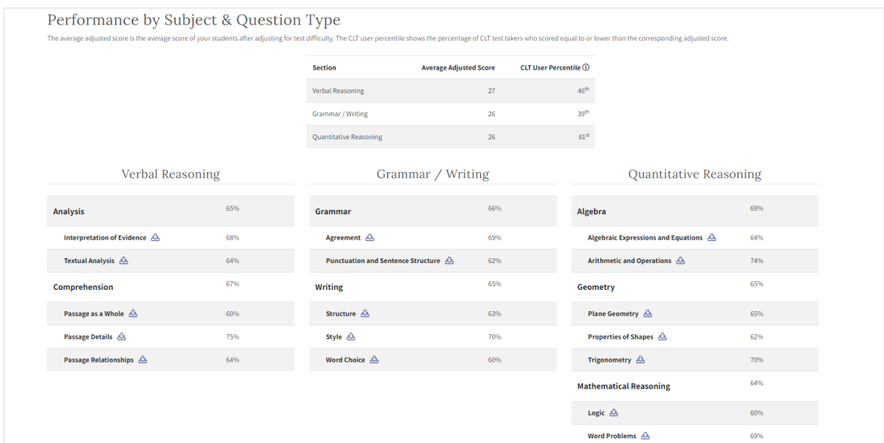
Table 6.3.1 CLT Sections, Domains, and Subdomains

SUBJECT SECTION	VERBAL REASONING		GRAMMAR/WRITING		QUANTITATIVE REASONING		
Domain	Analysis	Comprehension	Grammar	Writing	Algebra	Geometry	Mathematical Reasoning
Subdomain	Interpretation of Evidence	Passage as a Whole	Agreement	Structure	Algebraic Expressions and Equations	Coordinate Geometry	Logic
	Textual Analysis	Passage Details	Punctuation and Sentence Structure	Style	Arithmetic and Operations	Properties of Shapes	Word Problems
		Passage Relationships		Word Choice		Trigonometry	

Analytics are delivered to schools on the test administrator level, and student-level. For each level, a percent correct metric is given for each domain and subdomain. At the school-level, this percent correct metric displays the average percentage of questions the students at that school got correct within the specified category, for the specified test.

School administrators can see the top and bottom four domain-subdomain pairings (in terms of performance), as well as a breakdown of how the school performed on each section, domain, and subdomain, as pictured below.





7. EQUATING, SCALING, AND SCORING

7.1 Introduction

Each academic year, CLT develops multiple test forms with unique items to ensure test security. This prevents the items from being shared or remembered from previous test attempts. Given that students take different versions of the test, it is crucial that scores are comparable across forms. For example, two students who took different forms but have the same ability should receive the same score regardless of the specific test form they were administered. However, if the items on the two forms vary in content and difficulty, and the forms are scored simply based on the number of correct responses, then the scores of these two students will not be comparable. Chapter 3 describes the automated test assembly procedure we use to ensure that students are administered test forms that are parallel in content and statistical specifications such as difficulty and measurement precision. In practice, it is difficult to build forms that are identical in difficulty, so further psychometric procedures are implemented during the scoring process to adjust for any potential form differences.

To ensure that CLT scores are consistent and comparable across administrations, CLT conducts a series of analyses using Item Response Theory (IRT). IRT consists of a family of models that model the probability of a correct response to an item as a function of the difference between student ability and item difficulty. Item and ability parameters obtained from different test forms are placed on the same scale through a calibration process that is described below (Kolen & Brennan, 2014). This calibration process enables us to both measure a student's ability independently of the items on a particular test form and measure an item's difficulty independently of the particular group of test-takers that took the item. Before reporting, ability estimates from the IRT model are transformed to scale scores to aid interpretability.

This chapter begins with an overview of the IRT model CLT uses, the Rasch model. Then, we explain the calibration methodology we use to construct a common scale for student abilities and item difficulties obtained from different test forms. Next, we describe the data cleaning process conducted before the IRT calibrations. Finally, we explain the process used to transform the ability estimates to the scale scores that are reported to students.

7.2 The Rasch Model

The Rasch model models the probability that a student will answer a given item correctly as a function of two parameters: the test-taker's ability and the item's difficulty. The more capable the student and the easier the item, the higher the odds that the student will get the item right. Mathematically, odds are defined as the ratio of probabilities. In this case, the odds refer to the ratio of the probability of answering an item correctly to the probability of answering it incorrectly. Taking the logarithm of the odds allows us to express the odds as a linear function of student ability and item difficulty (Equation 7.1):

$$\log\left(\frac{P_{ni}}{1-P_{ni}}\right) = \theta_n - b_i \tag{7.1}$$

where P_{ni} is the probability that test-taker n will answer item i correctly, θ_n is the ability of test-taker n, and b_i is the difficulty of item i.

Both the ability estimates and the difficulty estimates are on the log-odds scale, also called the logit scale. Consequently, item difficulty and test-taker ability can be directly compared to each other. In the Rasch model, item difficulty is defined as the ability level at which the probability of answering the item correctly is 50%. That is, students whose ability is higher than the item's difficulty will have greater than a 50% chance of answering the item correctly, whereas students whose ability is lower than the item's difficulty will have smaller than a 50% chance of answering the item correctly. Most observed logit values fall in the -3 to 3 range. The probability of answering an item correctly (P_{ni}) can be expressed directly as well (Equation 7.2):

$$P_{ni} = \frac{\exp(\theta_n - b_i)}{1 + \exp(\theta_n - b_i)} \tag{7.2}$$

The Rasch model assumes unidimensionality, local independence of responses conditional on the latent trait, and that the probability of a correct response follows the Rasch logistic function. Unidimensionality means that the items on the test measure only a single construct/ability. Chapter 9 shows that the three sections of the CLT measure three unidimensional constructs: Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning. Therefore, we use the Rasch model to calibrate the three sections separately. Local independence means that the responses of a student to any two items are uncorrelated after controlling for ability. For example, if the answer to an item is implied in a previous item, the two items will be correlated independently of ability. CLT carefully screens items to prevent such clueing, and we have conducted a study that ruled out the presence of passage-level local dependence (described later in this chapter). Finally, the fit of the Rasch logistic function to the response data is evaluated by fit statistics described below.

7.3 Concurrent Calibration with the Common-Item Nonequivalent Groups Design

DEFINING THE RASCH SCALE

The Rasch model has scale indeterminacy, which means that we can add an arbitrary constant to student abilities and item difficulties without changing the probability of a correct response to that item. Therefore, the model must be constrained in some way to allow parameter estimation. To define the Rasch scale, we constrain the mean of the ability distribution to zero. As a result, item difficulties are estimated relative to this zero point which reflects the mean of the student ability distribution.

Constraining the mean of student abilities determines the scale for the purposes of parameter estimation, but further steps need to be taken before we can compare ability and difficulty estimates across test forms. To illustrate why, suppose that group A takes form X and group B takes form Y, and suppose that the average ability of group A is higher than the average ability of group B. If the scale is identified by constraining the mean of test-taker abilities, the mean of the ability distribution will be set to zero for both groups even though the actual ability of group A is higher than the ability of group B. This means that even if two items in different forms end up with the same difficulty estimate, they cannot be considered equally difficult because the ability level "zero" that serves as the reference point for both analyses has a different meaning for different groups. Therefore, when Rasch analysis involves items from multiple test forms administered to groups that differ in ability, a calibration process is necessary to ensure that the logit values derived from the different forms are on the same scale and thus comparable. We describe this calibration process next.

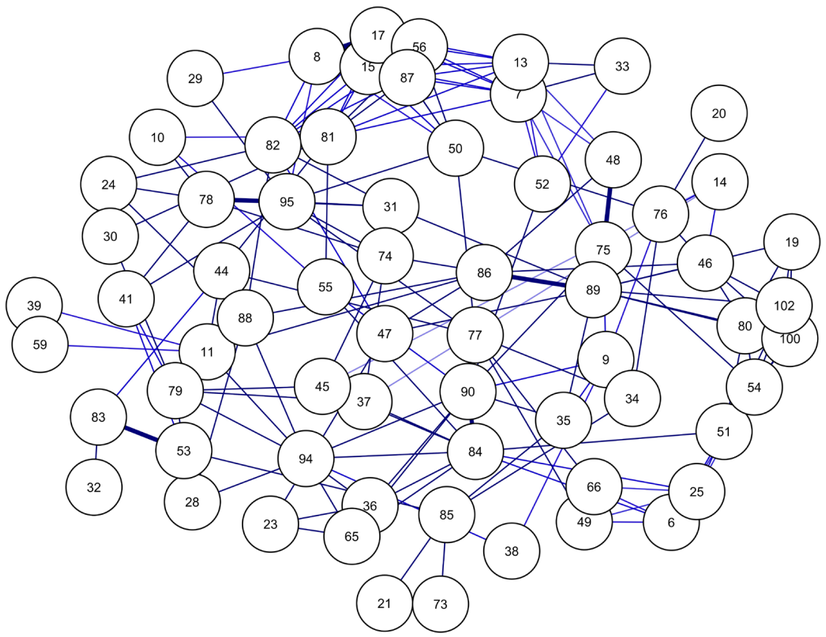
LINKING ITEM PARAMETERS FROM DIFFERENT FORMS

Items administered to different groups of students can be placed on the same scale if the test forms share a sufficient number of items that represent the characteristics of the test as a whole (Kolen & Brennan, 2014). This study design is known as the common-item nonequivalent group design, and the common items shared between the forms are referred to as anchor items (because they "anchor" the scale to a common metric). In the Verbal Reasoning and Grammar/Writing sections, the anchor items come from common passages. In the Quantitative Reasoning section, the anchor item sets are created from individual items. In our calibrations studies, we require a form to contain at least 10 anchor items to be calibrated. This represents 25% of the whole form (with 40 items), which is above the 20% minimum suggested by Kolen and Brennan (2004).

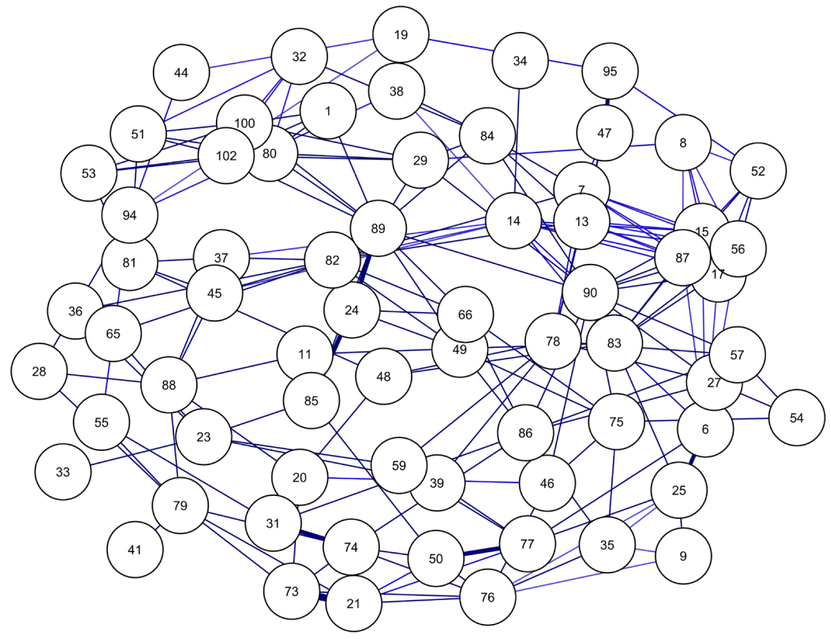
The links between different test forms can be conceptualized as a network in which the forms are the nodes and the common items are the edges connecting the nodes. Figure 7.1 presents the networks that show the common item structure of the CLT forms for each section. In the graphs, each node is a CLT form and the edges are the common items linking the forms. Thicker edges indicate a larger number of common items. Each number in the circles is a form ID. The networks were plotted with the **qgraph** package (Epskamp et al., 2012) in R (R Core Team, 2023).

Figure 7.1

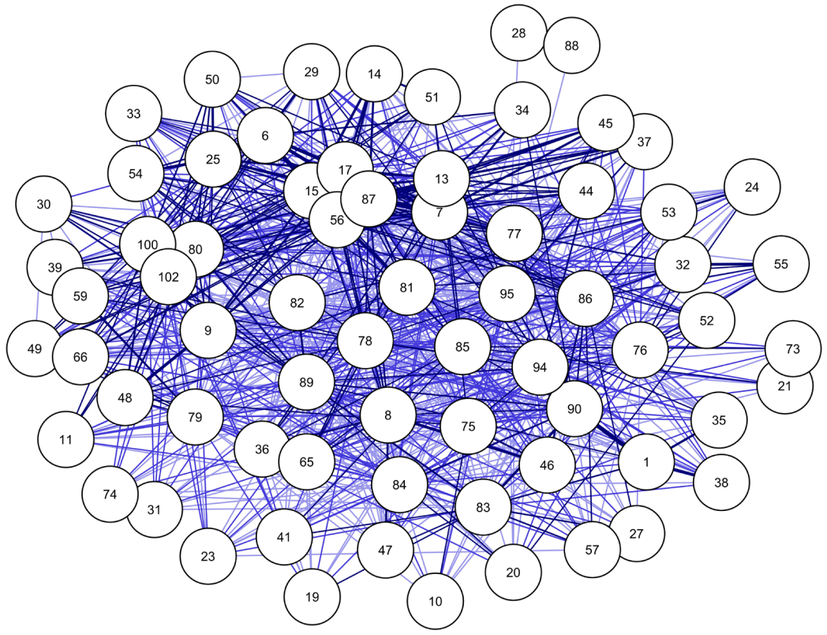
The Common Item Structure of CLT Forms



(a) The item network of Verbal Reasoning



(b) The item network of Grammar/Writing



(c) The item network of Quantitative Reasoning

The network of the Quantitative Reasoning section is denser because in addition to being connected by sets of items with a balanced content representation, Quantitative Reasoning forms can share small numbers of stand-alone items. On the other hand, Verbal Reasoning and Grammar/Writing items belong to passages and are used as intact sets of items. To prevent whole sets of items from being administered too often, Verbal Reasoning and Grammar/Writing passages are reused more sparingly. In the Quantitative Reasoning section, single items that are high quality can be reused more often.

CONCURRENT CALIBRATION

With this common item structure in place, we use the concurrent calibration methodology to estimate the difficulties of all the items on a common scale. Concurrent calibration involves stacking the response data of multiple test forms on top of each other to create a sparse matrix of student responses where most cells have missing values by design. This method is "concurrent" because the

combined data allows joint estimation of all item parameters in a single model while constraining anchor-item parameters to be equal across groups. This avoids separate post-calibration scale transformations such as estimating form-specific scales first and then linking them with methods like mean/mean or mean/sigma transformations. Thus, concurrent calibration is a rigorous way of linking the scales of separate forms that contain common items; it uses strong invariance as an explicit identification condition, places all forms on one scale in one step, and avoids a separate post-calibration transformation (von Davier & von Davier, 2004).

The concurrent calibration was run on forms administered between 2016 and 2023 to set a base Rasch scale. This base Rasch scale and the base items calibrated on that scale serve as the anchor point for all future forms and scores. All items written and administered after 2023 are brought onto the same scale by using the base items as anchors in a fixed-parameter calibration. For example, up to 10 items in each section of operational forms (25% of the form) are recently developed items with preliminary item statistics obtained from a machine learning model (see Chapter 3). After a new form has been administered, we estimate final item difficulties for these items using the base items on the form as anchors, and score students based on those final difficulty estimates. The base item parameters come from the concurrent calibration described here.

The calibrations were conducted using the **TAM** package (Robitzsch et al., 2025) in R (R Core Team, 2026) with marginal maximum likelihood (MML) estimation. As mentioned earlier in this chapter, the mean of the ability distribution was constrained to be zero. The analyses were conducted independently by two psychometricians, and all the item statistics were fully replicated.

ITEM DISCRIMINATION AND MODEL FIT

It is important that the items included in the calibrations, especially the anchor items used to link different forms, have high discrimination and fit the Rasch model well. To evaluate item discrimination, we use the point-biserial correlation r_{pb}. r_{pb} is the correlation between a binary variable such as item responses and a continuous variable such as total scores or latent ability. The value of r_{pb} ranges between -1 and 1. If an item has high discrimination, it is more likely to be answered correctly by students with high ability than low ability. Consequently, the responses to the item and the total scores obtained on the test will have a large, positive correlation. Conversely, if there is no relationship between student responses to an item and total scores, the point-biserial correlation will be close to 0. Sometimes, a negative point-biserial is observed, indicating that high ability students are less likely to answer the item correctly than low ability students. This may indicate an issue with the answer key and needs to be evaluated by test developers and content experts.

The point-biserial correlation is usually computed between the responses to an item and total

scores on the test, and the calculation of the total scores excludes the item being analyzed to prevent autocorrelation from inflating the statistic. However, in a concurrent setting where multiple forms are analyzed simultaneously, raw number correct scores on different forms are not directly comparable. Therefore, our calibration study used the point-biserial correlation between the responses to the item and the section θ estimated by the model (which is comparable across forms). As in the raw-score version of r_{pb}, θ is estimated without the item being analyzed. We used two cutoffs to qualify or disqualify a candidate anchor item: if the difficulty of the item was between -2 and 2, a r_{pb} threshold of < 0.15 was used to disqualify an item from being an anchor. If the difficulty of the item was smaller than -2 or larger than 2, a threshold of < 0.1 was used. This is because very easy or very hard items do not have as much response variability (most people get them right or wrong), which deflates r_{pb} even when the item is otherwise well functioning and has strong content validity.

The fit of the items to the Rasch model was examined using the outfit and infit mean squares (MSQ) (Wright & Masters, 1982). The outfit MSQ of an item is the average of its squared standardized residuals, which are the differences between the observed responses in the data and the response probabilities predicted by the model, divided by the modeled variance of the response (Equation 7.3).

$$OutfitMSQ_i = \frac{\sum\limits_n z_{ni}^2}{n} \tag{7.3}$$

where $z_{ni} = \frac{X_{ni} - E(X_{ni})}{\sqrt{Var(X_{ni})}}$, X_{ni} is the observed response of student n to item i, $E(X_{ni})$ is the expected response of student n to item i, and $Var(X_{ni}) = E(X_{ni})(1 - E(X_{ni}))$ is the variance of a student's response to an item. Outfit statistics are sensitive to outliers such as lucky guesses on hard questions by low-ability students or careless mistakes on easy questions by high-ability students. The infit MSQ, however, accounts for outliers by weighting the squared residuals by the proximity between an item's difficulty and a student's ability (Equation 7.4). For instance, for hard items, prediction errors for high-ability students are weighted more heavily than the prediction errors for low-ability students.

$$InfitMSQ_i = \frac{\sum\limits_n z_{ni}^2 \times Var(X_{ni})}{\sum\limits_n Var(X_{ni})} \tag{7.4}$$

where z_{ni} and $Var(X_{ni})$ are defined as above. Outfit and infit values have an expectation of 1. Values above 1 indicate model misfit, whereas values below 1 indicate model overfit, meaning that the responses are too predictable. Values above 1.5 are considered to be unproductive for measurement (Linacre, 2002). Therefore, items with an infit or outfit value above 1.5 were excluded from the calibrations.

Similarly, students whose responses are analyzed in the calibration can also fit or misfit the Rasch model, and the degree of misfit can be quantified using person infit and outfit statistics. High outfit values usually indicate guessing or careless mistakes, whereas high infit values indicate that the person's responses do not follow the Rasch model in a more substantive way. Students with an infit or outfit value above 1.5 were also excluded from the calibrations before diagnosing item misfit to prevent their responses from confounding item (mis)fit.

STABILITY OF ITEM PARAMETERS

Item difficulties obtained from the calibrations are used to develop forms (Chapter 3), score new forms, and calibrate newly developed items. In each of these uses, it is important to check that the parameters of calibrated items remain stable across time, forms, and populations to prevent drift in the common scale, biased parameter linking, and invalid score comparisons. Items may function differently across time and settings due to item drift, mode effects, or differential item functioning. Item drift refers to the fact that the difficulty of an item may change over time, mode effects mean that an item may have a different difficulty depending on the mode in which it was administered (i.e., online in-school, paper in-school, or remotely proctored), and DIF means the item has a different difficulty for different groups of people. DIF and mode analyses are presented in Chapter 9, so we do not elaborate on them here. To evaluate the stability of a given anchor item when scoring a new form, we re-estimate its parameter freely while keeping all the other items in the analysis fixed. Then, we compute the difference between the freely estimated parameter and the bank parameter used for anchoring. That is, for item j, we compute

$$\delta_j = \beta_j^{freed} - \beta_j^{bank}$$

Then, we transform this difference into a z-statistic using the pooled standard error of the freely estimated parameter and the bank parameter:

$$z_j = \frac{\beta_j^{freed} - \beta_j^{bank}}{\sqrt{SE_{freed}^2 + SE_{bank}^2}} \tag{7.5}$$

An item is flagged for drift if

$$|\delta_j| > 0.3$$

and

$$|z_j| \geq 1.96$$

where 1.96 is the critical value for a two-sided hypothesis test based on the standard-normal distribution. In other words, items are flagged for drift if the effect size of the drift is larger than 0.3 logits and if the drift is statistically significant. These items are then reviewed to investigate the cause of the drift. If a substantive cause is found, the difficulty parameter is re-estimated to reflect its current difficulty. If a substantive cause is not found, the bank parameter is retained to prevent the difficulty scale from drifting and potentially causing score inflation/deflation over time.

DATA CLEANING AND MISSING VALUES

Before running the calibration, we applied certain exclusion rules to ensure that the calibration samples were representative of the target population, the assumptions of the Rasch model were met, and the parameter estimates remained unbiased:

- Students who received accommodations were excluded.
- International students were excluded.
- Students with obvious guessing patterns were excluded (i.e., selecting the same answer choice for all the items and zig-zagging the entire test).
- Students who answered 10 or fewer items in a given section were excluded.
- Following Ludlow and O'Leary (1999), omitted items were treated as incorrect, not-reached items were treated as missing. Omitted items are items with missing responses followed by non-missing responses later in the test. Not-reached responses are missing response strings at the end of the test with no response following them. An omitted item means that the student read the item and decided not to answer, indicating that they did not know the answer. Not-reached items mean that the student never encountered the item, so we have no information on how they would have responded (N.B., in scoring, all missing responses are treated as incorrect. The differentiation between omitted and not-reached items only applies to item calibrations to prevent item position from arbitrarily inflating difficulty estimates.). As an example, consider the following response string in a hypothetical test with 11 items where 1 indicates a correct answer, 0 indicates an incorrect answer, and "x" indicates a missing response:

item_1	item_2	item_3	item_4	item_5	item_6	item_7	item_8	item_9	item_10	item_11
1	1	0	x	1	0	1	1	x	x	x

In this example, we assume that the student saw item 4, since they continued the test afterwards. That is an omitted item. On the other hand, items 9, 10, and 11 all have missing values, so it is more likely that the student never reached them (e.g., they ran out of time). Since the student never reached these questions, we do not know if they could have answered them correctly or not. Therefore, these questions are left as missing values during the calibrations. The distinction between omitted and not-reached items are only made during item calibration. When scoring students, all missing responses are treated as 0. For a detailed treatment of this approach with examples and an explanation of its advantages, we refer the reader to Ludlow and O'Leary (1999). The problems with treating not-reached items as incorrect are discussed in Mislevy and Wu (1996), Rose et al. (2017), Shin (2009), and Custer et al. (2012).

The sample sizes of the final calibration data were:

- 33,401 records in VR.
- 33,923 records in GW.
- 34,488 records in QR.

These records were not all from unique students and include multiple attempts from the same student. This was not expected to violate the local independence assumption of the Rasch model because a student rarely encounters an item again when they retake the test. In the rare occasions when a student answered the same item multiple times, only the first response was retained. We also fit a multilevel Rasch model to jointly account for the dependence structure between attempts, within passages (i.e., the testlet model), and across testing blocks (e.g., September 2020, December 2023, etc.). The correlation between the item parameters obtained from this model and the base Rasch model was 0.998 for VR and GW, and 1.000 for QR. The respective mean differences in item difficulties were -0.010 for VR, 0.037 for GW, and 0.024 for QR. This provided assurance that ignoring these dependence structures did not bias the item parameters.

CALIBRATION RESULTS

After data cleaning, we followed the concurrent calibration procedure described above and fit the Rasch model to each section. The result of this process is a calibrated item bank that has all the difficulty estimates on the same logit scale. Table 7.1 and Table 7.2 show the distribution of the difficulty (b) and the discrimination (r_{pb}) parameters of items that passed the item quality and fit checks, and are used operationally. Figures 7.2 and 7.3 show the respective frequency distributions.

Table 7.1

Distribution of the Item Difficulties

Section	N	Mean	SD	Min	Max	$b < -2.5$	$-2.5 \leq b \leq 2.5$	$b > 2.5$
VR	1443	-1.23	1.25	-5.39	4.85	15.66%	84.06%	0.28%
GW	1336	-1.37	1.43	-7.13	3.07	19.46%	80.31%	0.22%
QR	1580	-0.26	1.20	-4.99	3.61	3.67%	95.44%	0.89%

Note. The table shows the number of calibrated items and the mean, standard deviation, minimum, and the maximum of the item difficulties. The last three columns show the percentage of items within different difficulty ranges.

Table 7.2

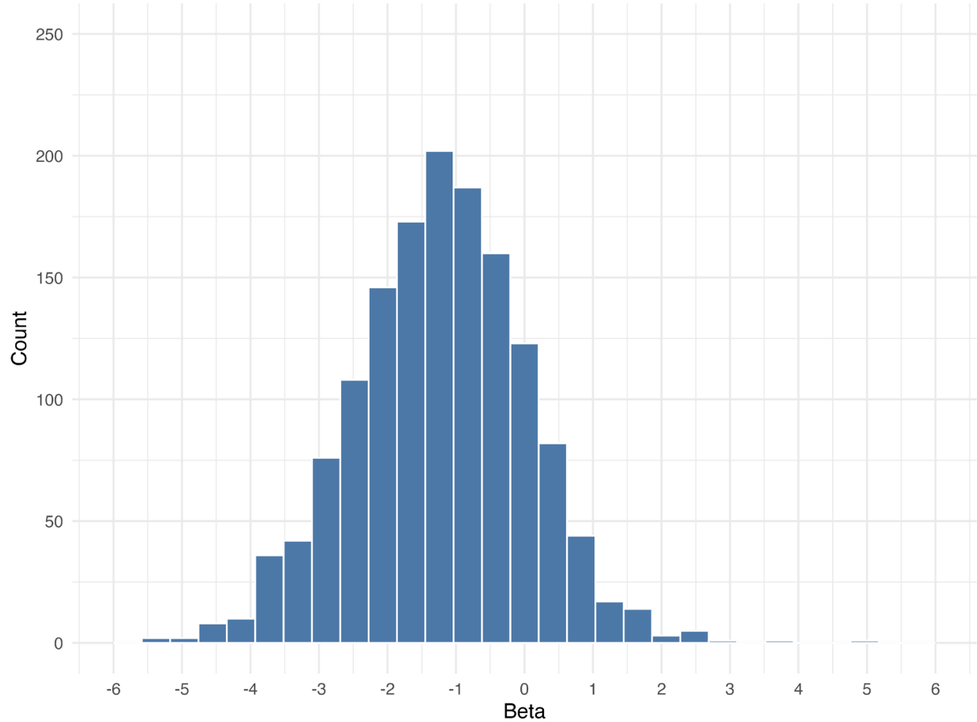
Distribution of the Item-Theta Correlations

Section	N	Mean	SD	Min	Max	$r_{pb} < 0.25$	$r_{pb} \geq 0.25$
VR	1443	0.30	0.08	0.10	0.59	26.68%	73.32%
GW	1336	0.30	0.09	0.10	0.58	31.89%	68.11%
QR	1580	0.31	0.09	0.10	0.56	27.59%	72.41%

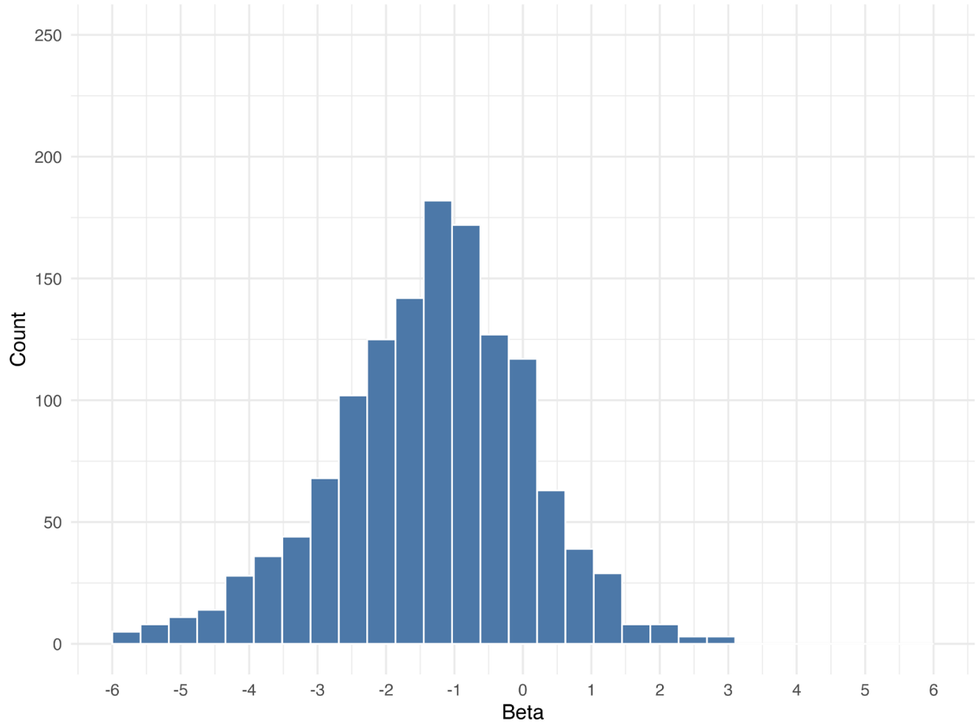
Note. The table shows the number of calibrated items and the mean, standard deviation, minimum, and the maximum of the item-theta correlations. Also, the last two columns show the percentage of items with an item-theta correlation in the given range.

Figure 7.2

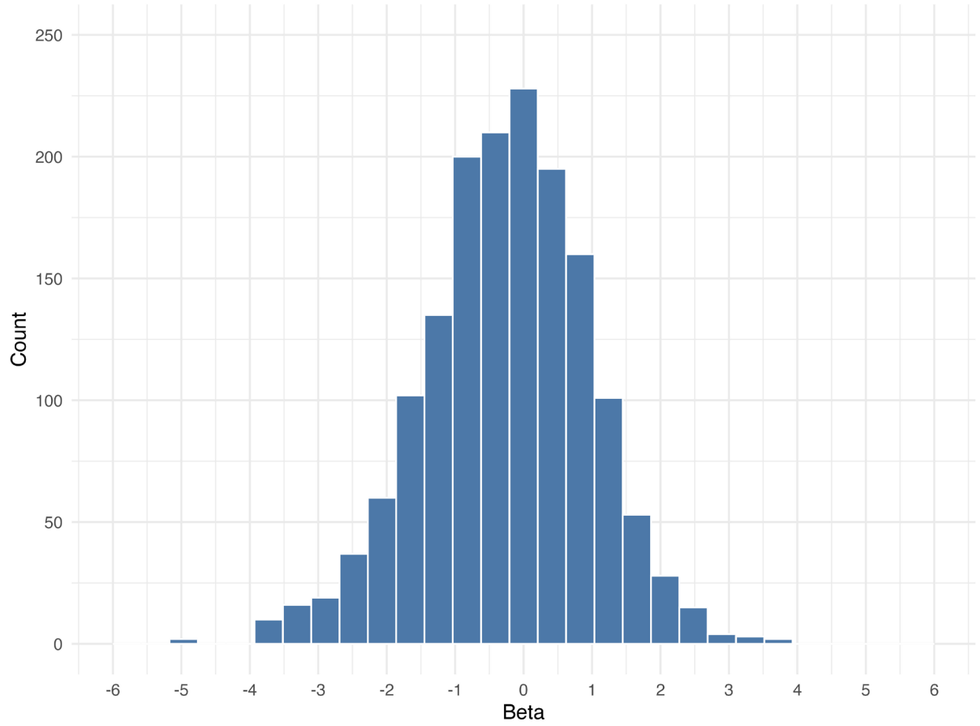
Frequency Distribution of Item Difficulties in Each Section



(a) Verbal Reasoning



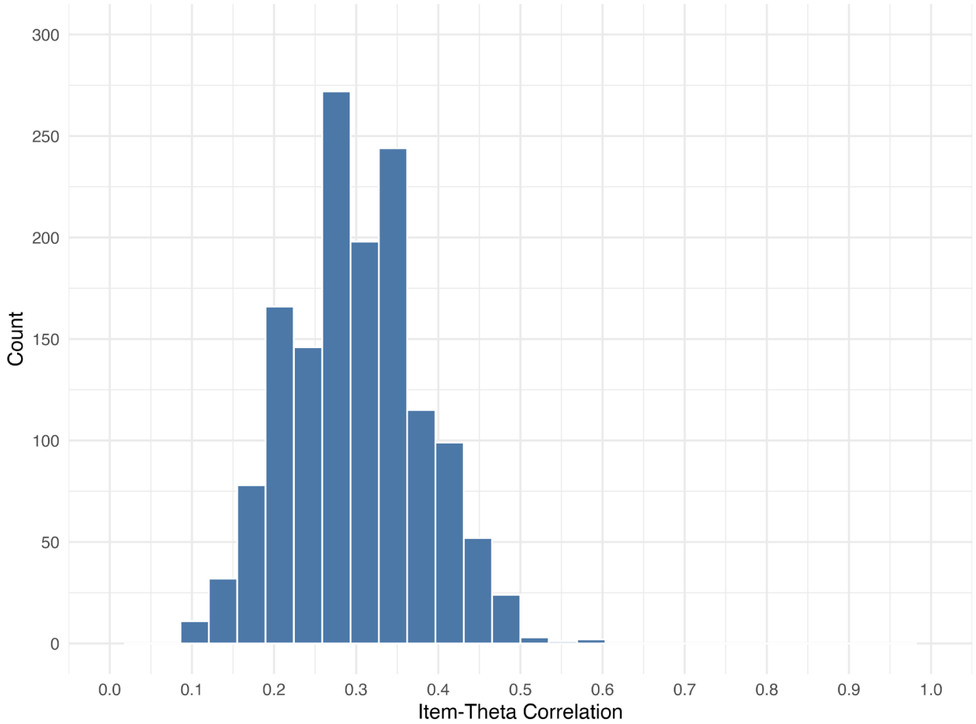
(b) Grammar/Writing



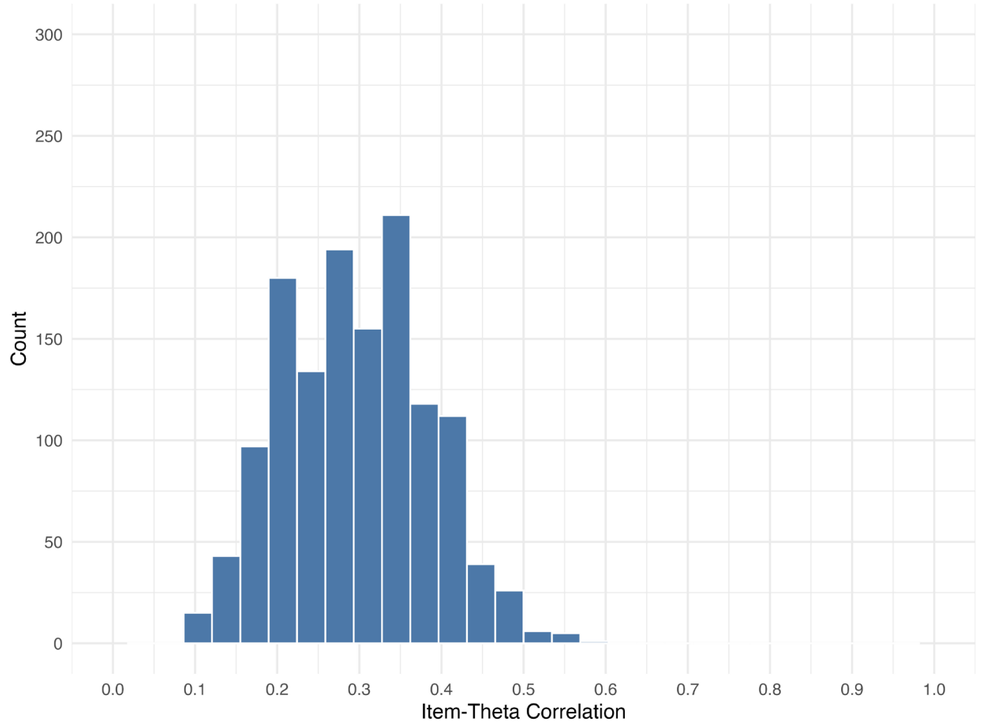
(c) Quantitative Reasoning

Figure 7.3

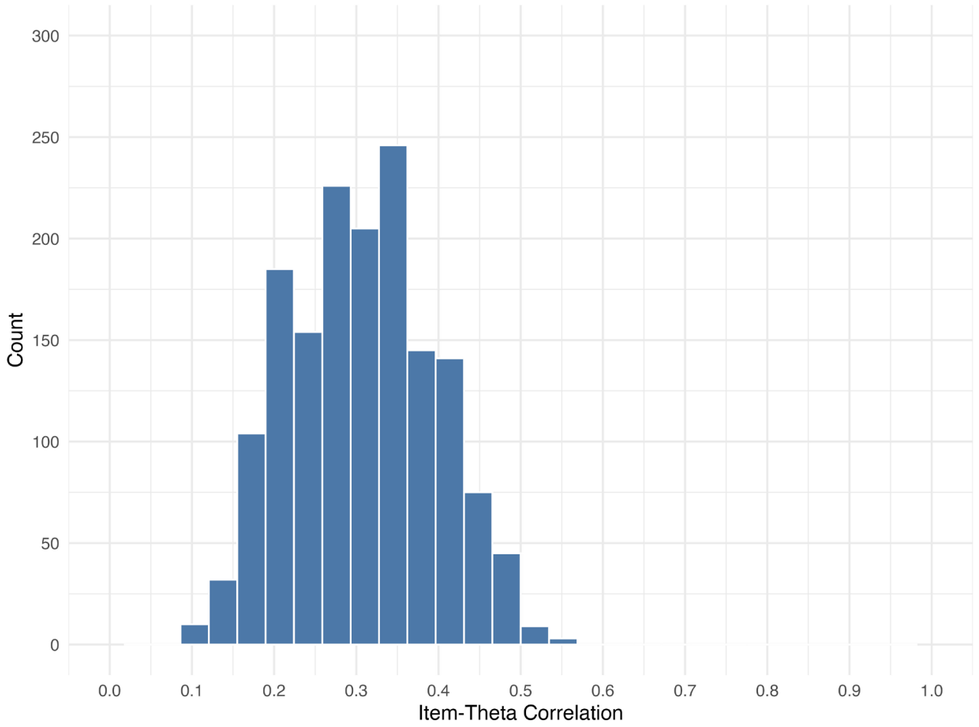
Frequency Distribution of Item-Theta Correlations in Each Section



(a) Verbal Reasoning



(b) Grammar/Writing



(c) Quantitative Reasoning

7.4 Scoring

Before we describe our scoring methodology, we state its goals:

1. The meaning of a CLT scale score must remain fixed across time. That is, our scoring methodology must prevent score inflation/deflation and preserve score comparability across years.
2. The reported scale scores, though derived from thetas, must remain on the 0-40 score scale that CLT users are familiar with and have historically used.

To accomplish these goals, we first set the reported score distribution of the 2016-2023 CLT user population as a reference distribution, in the same way their latent ability distribution was used to define the Rasch ability scale. Although this group was used to define the Rasch scale in the calibrations, they took the CLT before we used IRT for scoring, and their reported scores reflect raw-scores that were post-equated using Classical Test Theory (CTT). Having defined this reference

distribution, we operationalize the "meaning" of a score in the following way: if a hypothetical population that exactly matches the true ability distribution of the reference population were to take the CLT in, say, 2040, their scale score distribution should match the reported score distribution of the reference population as closely as possible.

Given that Rasch ability estimates from post-2023 administrations are already on the same scale, what's needed is a method to map each θ to a score point on the 0-40 scale for each section such that the meaning of the mapped score is the same as it was between 2016-2023. Linear transformations of θ do not work well in this context because, due to the S-shaped relationship between θ and the 0-40 raw score scale, a linear transformation of θ stretches the 0-40 scale and substantially changes the score distribution, especially at the tails.

Therefore, we used the base Rasch calibrations described earlier to retroactively estimate the thetas of the reference group, and then estimated an equipercentile link between their thetas and their reported scores in a single-group design. This equipercentile link was used to create a theta-to-scale score lookup table that covers theta values from -6 to 6 in increments of 0.0001. This amounts to having a direct lookup value for every possible theta rounded to 4 decimals, which is our operational rounding practice. Observed thetas outside the lookup range are truncated to ± 6.

To create the lookup table, we first compute theta percentiles in the reference group using the midpoint continuity correction described by Kolen and Brennan (2014) (explained below). We then linearly interpolate the percentiles of unobserved thetas across the theta grid using `approx()` in R (R Core Team, 2024). Finally, we use the inverse of the same percentile function along with linear interpolation to find the reported score in the same reference group that corresponds to a given theta percentile. This creates a map from each theta value to a scale score.

COMPUTATION OF PERCENTILES

Let F_θ be the cumulative distribution function (CDF) of θ in the reference group and F_S the CDF of reported scores. We define the quantile function of reported scores as the inverse of the CDF:

$$Q_S(p) = F_S^{-1}(p) \tag{7.6}$$

For a theta value, the corresponding equipercentile score is given by:

$$Y_S(\theta) = Q_S(F_\theta(\theta)) \tag{7.7}$$

We follow Kolen and Brennan (2014) in assuming that our scores — both theta and the reported

scores — are continuous values that have been discretized. So if x is a score value (either theta or a reported score), we can assume that half the people with a score of x are actually between x and $x + \epsilon$, and the other half are between x and $x - \epsilon$. Consequently, the corrected percentile of x is given by:

$$F_X(x) = \frac{\#(X < x) + 0.5 \cdot \#(X = x)}{N} \tag{7.8}$$

where # denotes counts. We also apply the following bounds so that percentiles stay within $[0.5/N, 1 - 0.5/N]$:

$$p^* = \min\left(\max\left(F_X(x), \frac{0.5}{N}\right), 1 - \frac{0.5}{N}\right) \tag{7.9}$$

Finally, we use linear interpolation via the `approx()` function in R to estimate percentiles for theta values not observed in the norm sample.

INVERTING OF PERCENTILES

Suppose we have a set of reported scores sorted in increasing order: $y_{(1)} \leq \cdots \leq y_{(k)} \leq \cdots \leq y_{(n)}$. Since the order statistic k corresponds to the number of scores at or below $y_{(k)}$, by Equation 7.8, the percentile of $y = y_{(k)}$ is:

$$p = \frac{k - 1/2}{n} \tag{7.10}$$

But here we are given a percentile p and want to invert it to find the position of $y_{(k)}$. Let h be our estimate of that position. Then by inverting Equation 7.10, we get:

$$h = pn + 1/2 \tag{7.11}$$

Notice that unlike k, h is a continuous value. If we define j as the largest integer smaller than h (i.e., the floor of h, $\lfloor h \rfloor$), then j is the integer part of h, and $w = h - j$ is the fractional part of h. That fractional part w tells us where h lands between j and $j + 1$. Thus, we can use it as a weight to do a linear interpolation when the scale score that corresponds to a given theta lies between two observed values y_j and y_{j+1}:

$$Q_S(p) = (1 - w)y_{(j)} + w\, y_{(j+1)} \tag{7.12}$$

The final scale score is $Q_S(p)$ rounded to the nearest integer.

SCORING A NEW TEST FORM

After an administration, scoring a form involves several steps. As Chapter 3 explains, up to 10 items in a form (25%) represent new content with preliminary item statistics obtained from a machine learning model. The first step in the scoring process is to estimate final item difficulties for these items by anchoring the remaining items to their bank difficulties. Even if the final estimates for these

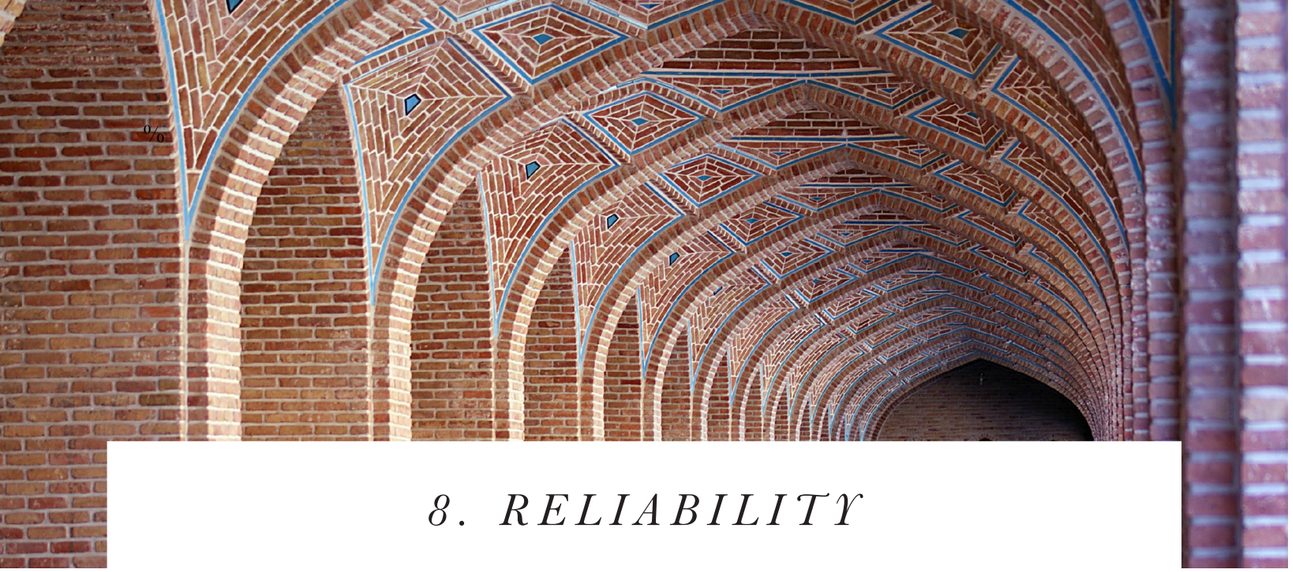
new items deviate from their preliminary estimates, the Rasch model allows us to calibrate them on the same difficulty scale as the other items before scoring by fixing the anchor items to their bank parameters.

The calibration of new items also involves item analyses to detect any unexpected fit, content, or drift issues in both the bank items and the new items (though drift analyses do not apply to new items). Specifically, we use the point-biserial correlation, infit/outfit statistics, and the drift detection procedure described earlier in this chapter to flag items with potential content or psychometric issues. Flagged items are reviewed by content experts to assess if there is a substantive reason that can explain their fit/drift statistics. Items that pass the review for both content and parameter stability are scored using their bank parameters. Items that pass the content review for fit but show substantive drift are freed, their difficulties are re-estimated using the remaining items as anchors, and the updated parameter is used to score the test. Items that fail the content review are excluded from scoring.

Excluding an item is a rare occurrence, and when it does happen, the section can still be scored on the same 0-40 scale because student scores are based on Rasch theta estimates from the remaining calibrated items, rather than on raw scores alone. As explained at the beginning of this chapter, Rasch thetas are on the same scale regardless of the specific set of scored items. In no case have we had to remove enough items from scoring to cause a material change in the blueprint; due to our experience in developing and targeting the difficulty of field test items for our CLT product, our maximum loss rate for field test items per administration is 1/10 (10%) field test items per section. These low loss rates allow us to consistently maintain the blueprint, difficulty distributions, and reliability for all of our operational forms, while ensuring our item pools remain refreshed.

Once the item reviews have been completed, **TAM** (Robitzsch et al., 2025) is used to obtain a raw-to-theta conversion table for each section by fixing the difficulty of each item to its known/estimated value. These conversion tables map each raw score on the test to a θ value rounded to 4 decimals. Then, the theta-to-scale score look-up tables described earlier in this chapter are used to map each theta to a scale score. Perfect raw scores are converted to the highest obtainable scale score (HOSS) (i.e., 40), and zero raw scores are converted to the lowest obtainable scale score (LOSS) (i.e., 0). After the scale scores are calculated for each section, they are summed to obtain a total CLT scale score. These total and section scale scores are then reported to students.

All the scoring steps described above are carried out independently by two psychometricians, and both the reported scores and the item parameters are fully replicated for each test.



8. RELIABILITY

8.1 Introduction

The reliability of test scores pertains to the precision and consistency of the scores a test produces. Validity, on the other hand, addresses the degree to which a test measures the construct it was designed to measure. Test scores must be reliable to be valid, but they do not have to be valid to be reliable (i.e., a test could reliably measure a construct that is different from the one it was designed to measure). Reliability is the focus of this chapter; validity is discussed in Chapter 9.

Test scores can be influenced by errors stemming from various random factors. For instance, a student might perform sub-optimally due to poor sleep the previous night or score higher than their true ability would suggest due to sheer luck (e.g., guessing correctly on items). Classical Test Theory (CTT) formalizes this concept by separating test scores into two components: a true score and an error component (Equation 8.1):

$$X = T + E \tag{8.1}$$

where X represents the observed score (number of correct answers), T signifies the true score, and E denotes the error. A larger error implies larger variability of observed scores around the true score. The standard error of measurement (SEM) corresponds to the standard deviation of the observed scores around the true score. In other words, SEM quantifies the spread of the error term.

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) state that in addition to SEM, conditional standard errors of measurement (CSEM) are reported for each score point. This is because measurement precision is not constant across the score scale, and modern measurement theories such as IRT enable the estimation of standard errors for specific ability levels. Therefore, we use CTT to report an overall reliability coefficient at the section level and IRT to report measurement precision at different scale scores for each section. Specifically, Sections 8.2-8.3 use Cronbach's alpha (Cronbach, 1951) to estimate reliability and SEM at the test level whereas Section 8.4 shows CSEM for various ranges of scale scores.

Reliability is reported for the two CLT administrations in 2026 (so far): January 2026 and February 2026. The January administration included in-school online and remotely-proctored tests which used the same form. The February administration involved all three modes in which the CLT is administered: in-school online, in-school paper, and remotely-proctored. The Quantitative Reasoning section of the paper form differed from the online form by one item, but otherwise the forms included the same items. This allows us to present reliability results for each mode separately in addition to the overall reliability of the scores from the February administration.

Finally, we note that the CLT user population has undergone significant changes since September 2023, when the state of Florida passed legislation that included the CLT among assessments that can be used to attain both state scholarships and high school graduation equivalence. Consequently, a large number of students who took the CLT in the last three years were not college-bound and used the test for high school graduation. Given that this population is a substantial part of CLT test takers, the results presented in this chapter include them as well as the historical college-bound CLT test takers. However, Chapter 10 presents reliability and validity results specifically for the college-bound population for interested readers.

8.2 Quantifying Reliability

Reliability can be quantified as the proportion of observed score variance that is due to true score variance (Harvill, 1991):

$$r_{XX'} = \frac{s_T^2}{s_X^2} \tag{8.2}$$

where $r_{XX'}$ denotes the reliability of the test scores, s_T^2 is the variance of true scores, and s_X^2 is the variance of observed scores. This expression can be re-written as

$$r_{XX'} = 1 - \frac{s_E^2}{s_X^2} \tag{8.3}$$

where s_E^2 is the error variance. Thus, the error variance becomes $s_E^2 = s_X^2(1 - r_{XX'})$ and the SEM is:

$$SEM = s_E = \sqrt{s_X^2(1 - r_{XX'})} = s_X\sqrt{(1 - r_{XX'})} \tag{8.4}$$

The most commonly used reliability coefficient is Cronbach's alpha, which measures the internal consistency of a test by examining the covariance between the items (Tavakol & Dennick, 2011). Internal consistency is the degree to which the items in a test measure the same latent construct and are related to each other. The formula for Cronbach's alpha is given in Equation 8.5 (Bland & Altman, 1997):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_X^2}\right) \tag{8.5}$$

where k is the number of items in the test, s_i^2 is the variance of item i, and s_X^2 is the variance of the total number-correct scores. Cronbach's alpha is affected not just by the variances and covariances of items and total scores but by test length as well; adding similar items to a test form will increase alpha. Cronbach's alpha takes values between 0 and 1, and the psychometric literature has suggested acceptable values that range from 0.70 to 0.95 with no consensus on what value of alpha is "good" or "high" (Taber, 2018; Tavakol & Dennick, 2011). That said, state education departments often require a reliability coefficient of 0.80 or above, and consider 0.80-0.89 as high (Texas Education Agency, 2022; Florida Department of Education, 2023).

Cronbach's alpha is a sample-dependent statistic, meaning that it does not estimate a test's reliability in general but the reliability of the scores obtained by a specific sample of examinees (Graham, 2006). Also, Cronbach's alpha assumes unidimensionality, which means that the items must measure a single latent construct (Cho & Kim, 2014). Given that the CLT measures the three constructs of Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning (Chapter 9), we report Cronbach's alpha for each CLT section separately. Cronbach's alpha was calculated using the **psych** package (Revelle, 2026) in the R programming language (R Core Team, 2023).

8.3 The Reliability and SEM of January and February 2026 Administrations

Table 8.1 presents the Cronbach's alpha and SEM for the January and February 2026 online administrations. Table 8.2 shows the reliability of the February form in each test mode. Each section of each administration met or exceeded an alpha of 0.80, most of them being significantly higher.

Table 8.1

The Reliability of Each Section of the January-February 2026 Online Administrations

Section	January 2026		February 2026	
	α	SEM	α	SEM
Verbal Reasoning	0.87	2.83	0.88	2.77
Grammar/Writing	0.88	2.78	0.83	2.73
Quantitative Reasoning	0.87	2.75	0.83	2.67

Table 8.2

The Reliability of Each Section by Test Mode (February 2026 Administration)

Section	In-School Online		Remotely-Proctored		In-School Paper	
	α	SEM	α	SEM	α	SEM
Verbal Reasoning	0.87	2.78	0.89	2.73	0.92	2.70
Grammar/Writing	0.81	2.72	0.85	2.77	0.90	2.72
Quantitative Reasoning	0.80	2.65	0.87	2.76	0.86	2.69

8.4 Conditional Standard Errors of Measurement (CSEM)

Chapter 7 explained that CLT's scoring methodology involves two steps: 1) the estimation of Rasch thetas from student responses; 2) mapping of thetas to scale scores using an equipercentile link between the thetas and reported scores of a reference population. As a result, the standard error of a scale score depends on two sources of uncertainty: 1) the standard error of the theta that maps to that scale score; 2) the standard error of the equipercentile link between theta and the scale score. Next, we describe how these two standard errors are estimated and combined to estimate the standard error of scale scores. Readers who want to skip to the results should refer to Tables 8.3-8.8, which show the median CSEM of various score intervals for each section of the January and February 2026 administrations.

STANDARD ERRORS OF RASCH THETAS

The precision of a Rasch theta depends on both the number of items used for measurement and the match between item difficulty and person ability. This makes intuitive sense: if an item is too easy or too hard for a student, the student will almost certainly get the item right/wrong, and their response will not provide much information about where their ability sits. In contrast, if the difficulty of the item is right at the student's ability, then the student will have about a 50% chance of getting the item right (see Chapter 7), which gives strong signal that their ability is close to that item's difficulty. The item information function (IIF) computes the information in a single item response about a given ability level, and is given by $p_{ij}(1-p_{ij})$ for the Rasch model, where p_{ij} is the probability that student i will answer item j correctly. IIF is maximized when $p_{ij} = 0.5$, which means the item's difficulty matches the student's ability perfectly. The test information function (TIF) is the sum of IIFs, and gives the total information the set of items provide about a given ability (Equation 8.6):

$$I(\theta_i) = \sum_{j=1}^{k} p_{ij}(1-p_{ij}) \tag{8.6}$$

where k is the total number of items. The standard error of theta, or the CSEM on the theta scale, is the square root of the inverse TIF (Equation 8.7):

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{8.7}$$

where $I(\theta)$ is the test information function as defined above.

CONTRIBUTION OF THETA UNCERTAINTY TO THE CSEM OF SCALE SCORES

The standard error of a Rasch theta does not translate linearly to the standard error of a scale score because the theta-to-scale-score link is discrete rather than continuous (because the scale scores are integers). As a result, multiple thetas correspond to the same scale score. Consequently, a small change in theta may leave the reported score unchanged if the theta remains within the same score interval, but it may also move the student across a score boundary and change the reported score. Therefore, we did not use a delta-method approximation to translate the standard error of theta to the standard error of a scale score. Instead, we treated each reported score as corresponding to an interval on the theta scale, and used the standard error of theta to calculate the probability that a student's true theta would fall into each interval that maps to a given scale score.

Let $\hat{\theta}_i$ denote the Rasch theta estimate for student i, and let $SE(\hat{\theta}_i)$ denote the standard error of that theta from Equation 8.7. Let $s_1, s_2, ..., s_m$ be the distinct reported scores in the equipercentile lookup table. For each reported score s_k, define the theta interval $[a_k, b_k)$ that maps to that score, where a_k is the smallest theta value that maps to s_k and b_k is the smallest theta value that maps to the next reported score. For the lowest reported score, we set $a_1 = -\infty$, and for the highest reported score, we set $b_m = \infty$.

We then approximated the uncertainty around $\hat{\theta}_i$ with a normal distribution:

$$\Theta_i \sim \mathcal{N}\left(\hat{\theta}_i, SE(\hat{\theta}_i)^2\right) \tag{8.8}$$

Under this approximation, the probability that student i maps to reported score s_k under the theta uncertainty is:

$$\pi_{ik} = P(a_k \leq \Theta_i < b_k) = \Phi\left(\frac{b_k - \hat{\theta}_i}{SE(\hat{\theta}_i)}\right) - \Phi\left(\frac{a_k - \hat{\theta}_i}{SE(\hat{\theta}_i)}\right) \tag{8.9}$$

where Φ is the standard normal cumulative distribution function. The expected reported score for student i is then:

$$\mu_i = \sum_{k=1}^{m} s_k \pi_{ik} \tag{8.10}$$

and the variance in reported scores induced by theta uncertainty is:

$$\mathsf{Var}_\theta(S_i) = \sum_{k=1}^{m} (s_k - \mu_i)^2 \pi_{ik} \tag{8.11}$$

Equation 8.11 gives the variance contribution of theta uncertainty to the CSEM of the reported score. This component reflects the fact that even if the equipercentile link were known perfectly, uncertainty in $\hat{\theta}_i$ would still create uncertainty in which scale score interval the student belongs to.

STANDARD ERRORS OF THE EQUIPERCENTILE LINK BETWEEN THETA AND SCALE SCORES

We used bootstrapping to estimate the standard errors of the equipercentile link between theta and reported scores in the reference group. This involved the following steps (please see Chapter 7 for details of the scoring methodology if the below procedure seems opaque):

1. For each test administration A that went into the estimation of the equipercentile link, resample with replacement N_A students where N_A is the sample size of administration A. Doing the resampling by administration ensured that the sample composition used to estimate standard errors of linking matched the sample composition used to estimate the equipercentile link itself.
2. Re-estimate the equipercentile link between each theta point in the theta grid and the reported scores in the re-sampled reference population.
3. Repeat steps 1 and 2 $B = 2000$ times.
4. For each theta point in the theta grid, take the standard deviation of the reported scores to which that theta mapped across the 2000 replications. This is the standard error of the equipercentile link.

More formally, let $L_b(\theta_g)$ denote the reported score assigned to theta grid point θ_g in bootstrap replication b, where $b = 1, 2, ..., B$ and $B = 2000$. Then the standard error of the equipercentile link at θ_g is:

$$SE_{link}(\theta_g) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(L_b(\theta_g) - \bar{L}(\theta_g) \right)^2} \tag{8.12}$$

where $\bar{L}(\theta_g)$ is the mean reported score assigned to θ_g across bootstrap replications.

CONDITIONAL STANDARD ERRORS OF SCALE SCORES

The final CSEM of a reported scale score combines the two sources of uncertainty described above: 1) uncertainty in the student's theta estimate, which induces uncertainty in the reported score even if the theta-to-score link were fixed; and 2) uncertainty in the equipercentile link itself. Let $\theta_g(i)$ denote the theta grid point matched to student i's estimated theta. Then the final CSEM of the reported score is:

$$CSEM(S_i) = \sqrt{\mathsf{Var}_\theta(S_i) + SE_{link}(\theta_g(i))^2} \tag{8.13}$$

where $\mathsf{Var}_\theta(S_i)$ is the variance from Equation 8.11. Thus, the first term captures the contribution of theta uncertainty under a fixed equipercentile link, and the second term captures the uncertainty in the equipercentile link itself.

Tables 8.3-8.8 show the CSEM of scale scores which were grouped into 5-point intervals for presentation. The reported CSEM is the median CSEM in the interval. We note that even though the scale scores are derived from thetas, the conditional standard errors do not follow the typical pattern of theta standard errors, which are smallest in the middle of the theta range and largest at the extremes. Rather, the middle score intervals often have higher errors than the tails, though the difference is small and the errors are fairly stable across the score range. This is because the scale score CSEM is a combination of theta and equipercentile linking errors; the equipercentile linking function is steepest in the middle of the score scale, so adjacent score points correspond to narrower theta intervals in the middle than in the tails. As a result, even though theta standard errors are smaller in the middle, they may translate to greater uncertainty in the scale score there because a given amount of theta uncertainty can be more likely to cross a score boundary.

Table 8.3

The CSEM of the January 2026 Verbal Reasoning Scale Scores

Score Interval	Median CSEM
10-15	2.64
15-20	2.68
20-25	2.65
25-30	2.41
30-35	2.05
35-40	1.69

Table 8.4

The CSEM of the January 2026 Grammar/Writing Scale Scores

Score Interval	Median CSEM
10-15	2.41
15-20	2.60
20-25	2.68
25-30	2.67
30-35	2.45
35-40	2.21

Table 8.5

The CSEM of the January 2026 Quantitative Reasoning Scale Scores

Score Interval	Median CSEM
10-15	2.50
15-20	2.70
20-25	2.81
25-30	2.79
30-35	2.58
35-40	2.34

Table 8.6

The CSEM of the February 2026 Verbal Reasoning Scale Scores

Score Interval	Median CSEM
10-15	2.43
15-20	2.76
20-25	2.96
25-30	2.89
30-35	2.68
35-40	2.51

Table 8.7

The CSEM of the February 2026 Grammar/Writing Scale Scores

Score Interval	Median CSEM
10-15	2.58
15-20	2.74
20-25	2.77
25-30	2.60
30-35	2.20
35-40	1.66

Table 8.8

The CSEM of the February 2026 Quantitative Reasoning Scale Scores

Score Interval	Median CSEM
10-15	2.52
15-20	2.70
20-25	2.79
25-30	2.79
30-35	2.54
35-40	2.25

9. VALIDITY

9.1 What is Validity?

The Standards for Educational and Psychological Testing define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p.11). In other words, validity relates to the interpretation of test scores, not the test itself. Testing organizations must provide evidence to validate the intended interpretations of the test scores. A valid test score interpretation is built upon high reliability. Thus, reliability is a prerequisite for validity. However, a reliable test may lead to invalid interpretations if the construct that it measures is different from the one it is intended to measure.

9.2 Sources of Validity Evidence

The Standards for Educational and Psychological Testing describe five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and consequences of testing (AERA, APA, & NCME, 2014). Evidence based on test content includes a description of the content domains that the test is intended to measure and an analysis of how the content of the test substantiates different aspects of the latent construct. Validity evidence based on content

was provided in the preceding chapters. This chapter provides validity evidence based on internal structure and relations to other variables.

Evidence based on internal structure analyzes the relationships between the items on a test to examine if the data support the hypothesized factorial structure of the latent construct that the test was designed to measure. For instance, items designed to measure mathematical reasoning should be strongly correlated with each other while showing a weaker relationship to other constructs such as reading comprehension. We use confirmatory factor analysis to evaluate the degree to which the CLT measures the three constructs represented by its sections: Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning.

The internal structure of a test is also related to measurement invariance in the sense that a test should measure the same construct in the same way for all subgroups of the population that uses the test. For example, if the test measures a different construct for males and females, the scores of males and females cannot be interpreted in the same way. Differential item functioning (DIF) is examined to evaluate if each item measures the same construct for each relevant subgroup. However, "the detection of DIF does not always indicate bias in an item; there needs to be a suitable, substantial explanation for the DIF to justify the conclusion that the item is biased" (AERA, APA, & NCME, 2014, p.51). Therefore, items that show DIF should be re-evaluated by content experts (Zieky, 2003).

Evidence based on relations to other variables includes convergent and discriminant evidence, test-criterion relationships, and validity generalization. Convergent evidence means that the scores obtained from the test that is being validated correlate strongly with scores obtained from other tests that measure a similar construct. Discriminant evidence means that the scores obtained from the test that is being validated correlate weakly with scores obtained from tests that measure a different construct. Test-criterion relationships concern the degree to which test scores predict an outcome of interest; when the criterion is measured at a later time, this evidence is referred to as predictive validity. Validity generalization refers to the degree to which such relationships generalize to new situations. This chapter provides predictive validity evidence based on the correlations between CLT scores and first-year college GPA, and convergent evidence based on the correlations between the CLT and the SAT®.

Factor analysis and DIF analyses were conducted again on the January and February 2026 forms. As in Chapter 8, which looked at CLT's reliability, we report CFA results for both the overall forms and each mode of the February 2026 administration separately. Likewise, DIF analyses are used to assess whether items function differently across modes after controlling for ability.

Also like Chapter 8, this chapter presents results for the entire group of students who took the

February and January forms, which includes both college-bound students and students who use the CLT for graduation equivalence. Chapter 10 provides validity and reliability evidence for the college-bound population specifically.

9.3 Validity Evidence Based on Internal Structure: Confirmatory Factor Analysis (CFA)

Psychological and cognitive constructs such as student ability are not directly observable. Therefore, they are called latent constructs. To measure a latent construct, observable behaviors that manifest the construct need to be identified. In standardized testing, the latent construct is the ability or the skill the test measures, and the observable behaviors are the students' responses to the test items. Thus, test items are also called indicators of the latent construct. The latent construct is assumed to underlie the indicators. For example, the responses of a student to questions of reading comprehension are modeled as a function of the student's ability to comprehend a text. In psychometrics, latent variables are studied using factor analytical methods, including exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA models do not have a priori assumptions about the factor structure that underlies the data. Instead, the number of factors and their structural relationships are uncovered from the data. When the researcher has a priori expectations about the factor structure, they can use CFA to test if the response data conform to their expectations.

The CLT has three sections designed to measure Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning. Therefore, we expect each section to be a separate unidimensional construct, which is also the assumption of Chapter 7 in scaling the three sections by separate Rasch models. To test that the three sections form three unidimensional constructs, we fit a separate one-factor confirmatory factor analysis (CFA) model to the dichotomous item responses within each section. Because the indicators/items are binary, they were treated as ordered categorical variables and the models were estimated using Weighted Least Squares Mean and Variance-adjusted (WLSMV). WLSMV views each observed 0/1 response as the categorized realization of an underlying continuous response variable. For identification, the variance of the latent factor was fixed to 1, and item loadings and thresholds were estimated relative to that standardized latent scale. The models were fit using the R package **lavaan** (Rosseel, 2012).

MODEL FIT

The fit of each model to the data was examined using the Root Mean Square Error of Approximation (RMSEA) (MacCallum et al., 1996), the Comparative Fit Index (CFI) (Bentler, 1990), and the Tucker-Lewis Index (TLI) (Tucker & Lewis, 1973). RMSEA is a measure of close fit and compares the model to a saturated model (i.e., a model that fits the data perfectly). CFI and TLI are incremental fit indices that compare the model to a baseline model in which the items are treated as unrelated to each other. In other words, RMSEA tells us the degree to which our model is worse than a perfect model while CFI and TLI tell us the degree to which our model is better than the worst possible model. RMSEA values below 0.05 indicate good fit, values between 0.05 and 0.08 indicate acceptable fit, and values above 0.08 indicate poor fit. CFI and TLI values above 0.90 indicate acceptable fit and values above 0.95 indicate good fit.

CFA RESULTS

Model fit for the one-factor section-level CFA models is summarized in Table 9.1. Verbal Reasoning showed consistently strong fit across administrations and modes. Its RMSEA values ranged from 0.026 to 0.032, and its CFI and TLI values ranged from 0.964 to 0.991 and from 0.962 to 0.991, respectively. These results support the interpretation of the Verbal Reasoning section as a clearly unidimensional construct. Grammar/Writing also fit the one-factor model well in general; RMSEA was consistently below 0.05, and CFI and TLI were above 0.90 in all cases except the February in-school online administration.

For Quantitative Reasoning, RMSEA indicated acceptable to good fit across all administrations and modes, ranging from 0.038 to 0.058. However, the incremental fit indices were mixed. The remotely-proctored February administration and the February paper administration met the CFI and TLI criterion for acceptable to good fit, with CFI/TLI values of 0.950/0.947 and 0.961/0.959, respectively. The January 2026 administration was borderline, with CFI = 0.901 but TLI = 0.895. By contrast, the combined February online sample and the February in-school online mode showed weaker incremental fit, with CFI/TLI values of 0.842/0.833 and 0.765/0.753, respectively. Chapter 10 shows that a unidimensional model fits the college-bound population's response data very well across the sections, forms, and modes.

Overall, the CFA results provide strong support for unidimensionality in Verbal Reasoning, generally supportive evidence for Grammar/Writing, and mixed evidence for Quantitative Reasoning, with RMSEA indicating acceptable to good fit but the incremental fit indices showing weaker support for a one-factor structure in some administrations.

Table 9.1

Fit of the One-Factor Section-Level CFA Models

Administration/Mode	Section	*N*	RMSEA	CFI	TLI
January 2026 Overall	Verbal Reasoning	6431	0.031	0.965	0.963
January 2026 Overall	Grammar/Writing	6431	0.036	0.959	0.957
January 2026 Overall	Quantitative Reasoning	6431	0.058	0.901	0.895
February 2026 Overall (Online Modes Combined)	Verbal Reasoning	19968	0.032	0.972	0.970
February 2026 Overall (Online Modes Combined)	Grammar/Writing	19968	0.041	0.911	0.906
February 2026 Overall (Online Modes Combined)	Quantitative Reasoning	19968	0.052	0.842	0.833
February 2026 In-School Online	Verbal Reasoning	16820	0.032	0.964	0.962
February 2026 In-School Online	Grammar/Writing	16820	0.040	0.891	0.885
February 2026 In-School Online	Quantitative Reasoning	16820	0.056	0.765	0.753
February 2026 Remote Proctored	Verbal Reasoning	3148	0.032	0.974	0.972
February 2026 Remote Proctored	Grammar/Writing	3148	0.037	0.941	0.938
February 2026 Remote Proctored	Quantitative Reasoning	3148	0.038	0.950	0.947
February 2026 Paper	Verbal Reasoning	1521	0.026	0.991	0.991
February 2026 Paper	Grammar/Writing	1521	0.026	0.986	0.986
February 2026 Paper	Quantitative Reasoning	1521	0.034	0.961	0.959

9.4 Validity Evidence Based on Internal Structure: Differential Item Functioning (DIF)

THE MANTEL-HAENSZEL PROCEDURE AND ETS CRITERIA

Differential item functioning shows the degree to which the difficulty of an item differs across demographic groups of interest after controlling for ability. A common way of assessing DIF is the Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959). In this procedure, test takers are divided into two groups: the reference group and the focal group. The performance of the reference group is taken as a reference point against which the performance of the focal group is compared. Then, groups are matched on a matching variable (typically raw scores). Once the groups are matched, the MH procedure calculates the conditional odds ratio of responding to an item correctly between the groups given the matched ability levels. The MH estimate of the conditional odds ratio is the aggregate of these conditional odds-ratios across the k levels of the matching variable (Equation 9.1) (Zwick, 2012):

$$\hat{\alpha}_{MH} = \frac{\sum_k \frac{N_{R1k} N_{F0k}}{N_k}}{\sum_k \frac{N_{R0k} N_{F1k}}{N_k}}, \qquad N_k = N_{R1k} + N_{R0k} + N_{F1k} + N_{F0k} \tag{9.1}$$

where N_{R1k} is the number of test takers in the reference group who answered correctly at the k-th ability level, N_{F1k} is the number of test takers in the focal group who answered correctly at the k-th ability level, N_{R0k} is the number of test takers in the reference group who answered incorrectly at the k-th ability level, N_{F0k} is the number of test takers in the focal group who answered incorrectly at the k-th ability level, and N_k is the total number of test takers at the k-th ability level. As a measure of effect size, Holland and Thayer (1985) developed the MH index to express MH on the Educational Testing Services (ETS) delta scale (Zwick, 2012):

$$\Delta_{MH} = -2.35 \ln \hat{\alpha}_{MH} \tag{9.2}$$

ETS places items in the categories of A, B, or C depending on the statistical significance as well as the effect size of DIF (Dorans & Holland, 1992). Category A means that DIF is negligible, category B means that DIF is moderate, and category C means that DIF is large (Magis et al., 2010). Categories B and C are further qualified by their signs: B+ and C+ indicate that the item favors the focal group whereas B- and C- indicate that the item favors the reference group (Zwick, 2012). An item's

category is determined by 1) the significance of two hypothesis tests; 2) the absolute value of Δ_{MH} (Dorans & Holland, 1992; Zwick, 2012). The first hypothesis test is of the null hypothesis that $\Delta_{MH} = 0$, which can be tested using the MH chi-square statistic (Dorans & Holland, 1992; Zwick, 2012):

$$MH_{\chi^2} = \frac{(|\sum_k N_{R1k} - \sum_k E(N_{R1k})| - \frac{1}{2})^2}{\sum_k Var(N_{R1k})} \tag{9.3}$$

The second hypothesis test is of the null hypothesis that $|\Delta_{MH}| \leq 1$ (or $-1 \leq \Delta_{MH} \leq 1$), which can be tested using Equation 9.4 (Zwick, 2012):

$$\frac{|\Delta_{MH}| - 1}{SE_{\Delta_{MH}}} > 1.645 \tag{9.4}$$

An item is in category A if Δ_{MH} is not significantly different from zero or $|\Delta_{MH}| < 1$, in category C if the absolute value of Δ_{MH} is both statistically significantly greater than 1 and larger than 1.5, and in category B if it does not meet the criteria for either category A or C.

DIF RESULTS

We analyzed DIF for gender, ethnicity, and test mode. For gender, males were the reference group and females the focal group. For DIF across ethnicities, White students were the reference group, and African-American and Hispanic/Latino students were the focal groups. For test modes, in-school online was the reference group and in-school paper and remotely-proctored were the focal groups. The MH statistics were calculated using the **difR** package (Magis et al., 2010), and the resulting MH estimates were then transformed into ETS delta values and classified according to the ETS significance and effect-size criteria described above.

Tables 9.2-9.4 summarize the distribution of items across the ETS categories for gender and ethnicity comparisons. Table 9.2 compares males and females, Table 9.3 compares White and Black students, and Table 9.4 compares White and Hispanic/Latino students. Across both administrations, the vast majority of items were classified in category A, indicating negligible DIF.

Gender DIF was minimal: all January items were classified in category A, and only one February Verbal Reasoning item was classified as B-. DIF for the White-Black comparison was also limited, with no flagged items in Verbal Reasoning in either form, two flagged items in January Grammar/Writing (one B+ and one B-), and three flagged items in January Quantitative Reasoning (one B+, one B-, and one C-). In February, the White-Black comparison showed only one B+ item in Quantitative Reasoning and no C+ or C- items.

The strongest ethnicity-based DIF was observed in the White-Hispanic/Latino comparison for the January form, where two Grammar/Writing items were classified as C- and three Quantitative Reasoning items were flagged (two B- and one C-). By contrast, all February items in the White-Hispanic/Latino comparison were classified in category A.

Mode DIF was also limited (Tables 9.5-9.6): the in-school online versus remotely-proctored comparison showed no flagged items in Grammar/Writing or Quantitative Reasoning and only two flagged Verbal Reasoning items, one B+ and one C+. The in-school online versus paper comparison showed no C+ or C- items and three B+ items across Grammar/Writing and Quantitative Reasoning.

Items in the C categories are reviewed by content experts to determine whether a substantive explanation exists and whether the item should be revised or removed. None of the items flagged in the January or February forms showed a content feature that warranted removal from scoring, though flagged items will continue to be monitored in future administrations.

Table 9.2

DIF Results for Males and Females

Form	Section	N_{Male}	N_{Female}	Total	A	B+	B-	C+	C-
January 2026	Verbal Reasoning	2660	3626	40	40	0	0	0	0
January 2026	Grammar/Writing	2660	3626	40	40	0	0	0	0
January 2026	Quantitative Reasoning	2660	3626	40	40	0	0	0	0
February 2026	Verbal Reasoning	9455	9899	40	39	0	1	0	0
February 2026	Grammar/Writing	9455	9899	40	40	0	0	0	0
February 2026	Quantitative Reasoning	9455	9899	40	40	0	0	0	0

Note. The plus (+) sign means that the item favors females, the minus (-) sign means that the item favors males.

Table 9.3

DIF Results for White and Black Students

Form	Section	N_{White}	N_{Black}	Total	A	B+	B-	C+	C-
January 2026	Verbal Reasoning	1867	1642	40	40	0	0	0	0
January 2026	Grammar/Writing	1867	1642	40	38	1	1	0	0
January 2026	Quantitative Reasoning	1867	1642	40	37	1	1	0	1
February 2026	Verbal Reasoning	4890	5224	40	40	0	0	0	0
February 2026	Grammar/Writing	4890	5224	40	40	0	0	0	0
February 2026	Quantitative Reasoning	4890	5224	40	39	1	0	0	0

Note. The plus (+) sign means that the item favors Black students, the minus (-) sign means that the item favors White students.

Table 9.4

DIF Results for White and Hispanic Students

Form	Section	N_{White}	$N_{Hispanic}$	Total	A	B+	B-	C+	C-
January 2026	Verbal Reasoning	1867	2066	40	40	0	0	0	0
January 2026	Grammar/Writing	1867	2066	40	38	0	0	0	2
January 2026	Quantitative Reasoning	1867	2066	40	37	0	2	0	1
February 2026	Verbal Reasoning	4890	7028	40	40	0	0	0	0
February 2026	Grammar/Writing	4890	7028	40	40	0	0	0	0
February 2026	Quantitative Reasoning	4890	7028	40	40	0	0	0	0

Note. The plus (+) sign means that the item favors Hispanic students, the minus (-) sign means that the item favors White students.

Table 9.5

DIF Results for In-School Online and Remotely-Proctored Students

Form	Section	N_{Online}	N_{RP}	Total	A	B+	B-	C+	C-
February 2026	Verbal Reasoning	16820	3148	40	38	1	0	1	0
February 2026	Grammar/Writing	16820	3148	40	40	0	0	0	0
February 2026	Quantitative Reasoning	16820	3148	40	40	0	0	0	0

Note. The plus (+) sign means that the item favors remotely proctored students, the minus (-) sign means that the item favors in-school online students.

Table 9.6

DIF Results for In-School Online and Paper Students

Form	Section	N_{Online}	N_{Paper}	Total	A	B+	B-	C+	C-
February 2026	Verbal Reasoning	16820	1521	40	40	0	0	0	0
February 2026	Grammar/Writing	16820	1521	40	39	1	0	0	0
February 2026	Quantitative Reasoning	16820	1521	39	37	2	0	0	0

Note. The plus (+) sign means that the item favors paper students, the minus (-) sign means that the item favors in-school online students. Quantitative Reasoning is based on 39 common items rather than 40 because one February 2026 paper item did not have an online counterpart and was excluded from the online-paper DIF analysis.

9.5 Predictive Validity: The Relationship Between CLT Scores and College GPA

We provide predictive validity evidence from two recent studies. First, (Welton, 2025) analyzed 235 students and found statistically significant positive correlations between all CLT scores and first-year GPA. The reported (raw) Pearson correlations were 0.37 for the CLT total score, 0.38 for Verbal Reasoning, 0.31 for Grammar/Writing, and 0.24 for Quantitative Reasoning. The study also reported a corrected correlation of 0.58 between CLT total score and first-year GPA

after adjusting for range restriction. In addition, hierarchical regression analyses indicated that demographic variables explained 7.3% of the variance in first-year GPA, and adding the CLT composite explained an additional 13.5%, suggesting that the CLT contributed meaningful predictive information beyond those background variables.

Second, a March 2026 study by CLT and Franciscan University (Classic Learning Initiatives & Franciscan University, 2026) analyzed data from 365 students and likewise found statistically significant positive correlations between CLT scores and freshman GPA. The reported raw Pearson correlations were 0.42 for the CLT total score, 0.27 for Verbal Reasoning, 0.35 for Grammar/Writing, and 0.36 for Quantitative Reasoning. After correcting for range restriction, the correlation between CLT total score and freshman GPA increased to 0.59. Regression analyses further showed that, after controlling for gender, race, religion, and high school GPA, the CLT sub-scores explained substantial additional variance in freshman GPA, with the full model accounting for 21.2% of the variance.

Taken together, these two studies indicate that CLT scores have a consistent positive relationship with first-year college GPA across distinct institutional settings. The raw correlation between CLT total score and first-year GPA ranged from 0.37 to 0.42, and the corrected correlation ranged from 0.58 to 0.59. These results are summarized in Tables 9.7 and 9.8. These findings are comparable to, and somewhat larger than, the predictive validity values currently reported by the SAT®; in its October 2024 national validity report (College Board, 2024), College Board reported a raw correlation of 0.32 and a corrected correlation of 0.53 between SAT® scores and first-year cumulative GPA.

Table 9.7

Predictive Validity Correlations with First-Year College GPA: Grove City College Study

Score	Raw	Corrected
Total	0.37	0.58
Verbal Reasoning	0.38	–
Grammar/Writing	0.31	–
Quantitative Reasoning	0.24	–

Note. The correlation values are from Welton (2025).

Table 9.8

Predictive Validity Correlations with First-Year College GPA: Franciscan University Study

Score	Raw	Corrected
Total	0.42	0.59
Verbal Reasoning	0.27	–
Grammar/Writing	0.35	–
Quantitative Reasoning	0.36	–

Note. The correlation values are from Classic Learning Initiatives & Franciscan University (2026).

9.6 Convergent Evidence: The Relationship Between the CLT and the SAT®

As discussed above, two tests that measure similar constructs are expected to have strong relationships. If one of the tests has already been accepted as a valid measure of the given construct, then the validity of a new test can be evaluated by analyzing the degree to which the scores it produces are correlated with the scores produced by the established test.

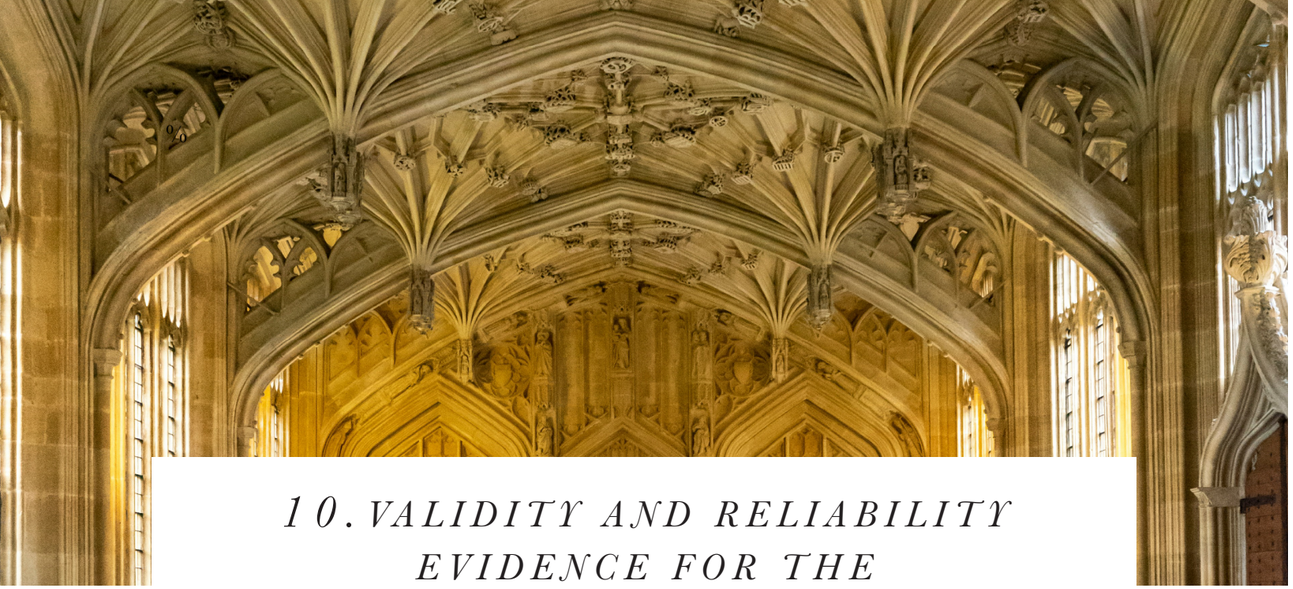
In April 2023, CLT conducted a concordance study between the CLT and the SAT® which also analyzed the correlation between the two tests (Classic Learning Initiatives, 2023a). Since the SAT® does not separate reading from writing and reports a combined Evidence Based Reading and Writing score, CLT Verbal Reasoning and Grammar/Writing scores were summed and correlated to the SAT® EBRW scores. Total scores were collected from 4,375 students who took both tests in the past. The sample size for section scores was 1,551. For more details of the study, readers are referred to the 2023 CLT & SAT® Concordance report and the Concordance Report Summary published on CLT's website.

The correlation between the total CLT scores and total SAT® scores was 0.89, the correlation between CLT VR+GW and SAT® EBRW was 0.90, and the correlation between CLT QR and SAT® Math was 0.87. These results are summarized in Table 9.9. Such high correlations provide strong evidence of convergent validity for the CLT. Moreover, the results of the content alignment study included in our concordance report emphasized that although the two tests differ in the specific types of texts (e.g., reading passages) they use, both tests measure the same latent constructs related to reading, writing, and mathematics.

Table 9.9

The Correlations Between the CLT and the SAT®

Section	N	Correlation
CLT Total – SAT® Total	4,375	0.89
CLT VR+GW – SAT® EBRW	1,551	0.90
CLT QR – SAT® Math	1,551	0.87



10. VALIDITY AND RELIABILITY EVIDENCE FOR THE COLLEGE-BOUND POPULATION

10.1 Background

In 2023, Florida enacted House Bill 1537, which authorized school districts to administer the CLT, allowed students to use concordant CLT scores for Bright Futures eligibility, and required the State Board of Education to establish CLT concordant and comparative scores for satisfying high school graduation assessment requirements (Florida Senate, 2023a, 2023b). These changes took effect in 2023, and in September 2023 the Florida Board of Governors separately approved the CLT as an admissions test for the State University System (Florida Department of Education, 2023; Florida Board of Governors, 2023). As a result of this legislation, Florida public school students have become the largest group among CLT test takers. However, the CLT score distribution of these students indicates that the majority of them take the test to satisfy graduation assessment requirements rather than for college admission. Specifically, the average CLT score has decreased from about 76 pre-September 2023 to 43 since September 2023.

Several of the reliability and validity analyses conducted in Chapters 9 and 10 were based on the full operational population of students who took the January and February 2026 CLT forms, which means that they reflect reliability and validity evidence for a sample that deviates from the typical target population of a college admissions test. Exceptions are conditional standard errors of measurement, which do not depend on specific administration data, predictive validity evidence which used a sample of students from Grove City College, and convergent validity evidence which was based on a diverse but predominantly college-bound sample. In contrast, analyses that relied on the January and February 2026 administrations, including internal consistency (i.e., Cronbach's alpha) and internal structure/unidimensionality (i.e., confirmatory factor analysis) were based on a mostly non-college-bound sample.

The purpose of this chapter is to conduct validity and reliability analyses for students who use the CLT for college admission. We have found that a practical exclusion rule that allows us to restrict the data set to college-bound students with high confidence while retaining a sufficient sample size for statistical analyses is to exclude students who are both from Florida and who are public school students. This is because the Florida legislation specifically applies to school districts, and due to the nature of the legislation, all Florida students who take the CLT to satisfy graduation requirements are public school students. We recognize that a minority of Florida public students who take the CLT also seek college admission, but when the nature of the legislation and the score distribution of the Florida public school students are considered together, it is very likely that the vast majority of the Florida public school students who take the CLT take it to satisfy graduation assessment requirements.

10.2 Reliability of the CLT for College-Bound Students

Table 10.1 presents Cronbach's alpha and SEM for the January and February 2026 online administrations after excluding Florida public school students. Table 10.2 shows the reliability of the February 2026 form by test mode for the same subgroup. Across the January and February online administrations, all alpha values were at or above 0.86. In the February mode-specific results, alpha values remained high for all the sections, with the lowest value being 0.83 for Quantitative Reasoning in the paper administration. Overall, these results indicate that the CLT continues to yield reliable section scores for college-bound samples.

Table 10.1

The Reliability of Each Section of the January-February 2026 Online Administrations for the College-Bound Population

Section	January 2026		February 2026	
	α	SEM	α	SEM
Verbal Reasoning	0.88	2.70	0.90	2.47
Grammar/Writing	0.90	2.60	0.87	2.62
Quantitative Reasoning	0.86	2.80	0.87	2.75

Table 10.2

The Reliability of Each Section by Test Mode for the College-Bound Population (February 2026 Administration)

Section	In-School Online		Remotely-Proctored		In-School Paper	
	α	SEM	α	SEM	α	SEM
Verbal Reasoning	0.91	2.46	0.89	2.47	0.89	2.50
Grammar/Writing	0.87	2.62	0.87	2.61	0.85	2.67
Quantitative Reasoning	0.88	2.73	0.86	2.77	0.83	2.76

10.3 Validity Evidence Based on Internal Structure: Confirmatory Factor Analysis

Using the same one-factor section-level CFA models described in Chapter 9, we re-estimated model fit after excluding Florida public school students. The fit results are summarized in Table 10.3. In contrast to the full operational population, model fit in the college-bound samples was uniformly strong across sections, administrations, and modes. RMSEA values ranged from 0.010 to 0.040, and all CFI and TLI values were at or above 0.961. Verbal Reasoning continued to show excellent fit across all administrations and modes. Grammar/Writing and Quantitative Reasoning also showed very strong fit in this subgroup, with no administration or mode showing the weaker incremental fit observed in Chapter 9 for the full operational population. Overall, these results provide strong evidence that the three CLT sections function as unidimensional constructs in the college-bound sample.

Table 10.3

Fit of the One-Factor CFA Models for the College-Bound Population

Administration/Mode	Section	*N*	RMSEA	CFI	TLI
January 2026 Overall	Verbal Reasoning	1130	0.040	0.963	0.961
January 2026 Overall	Grammar/Writing	1130	0.016	0.995	0.995
January 2026 Overall	Quantitative Reasoning	1130	0.023	0.984	0.983
February 2026 Overall (Online Modes Combined)	Verbal Reasoning	1583	0.021	0.991	0.991
February 2026 Overall (Online Modes Combined)	Grammar/Writing	1583	0.022	0.985	0.984
February 2026 Overall (Online Modes Combined)	Quantitative Reasoning	1583	0.017	0.990	0.990
February 2026 In-School Online	Verbal Reasoning	731	0.012	0.998	0.998
February 2026 In-School Online	Grammar/Writing	731	0.020	0.988	0.987
February 2026 In-School Online	Quantitative Reasoning	731	0.010	0.997	0.997
February 2026 Remote Proctored	Verbal Reasoning	852	0.020	0.990	0.990
February 2026 Remote Proctored	Grammar/Writing	852	0.016	0.991	0.991
February 2026 Remote Proctored	Quantitative Reasoning	852	0.012	0.994	0.994
February 2026 Paper	Verbal Reasoning	544	0.024	0.988	0.987
February 2026 Paper	Grammar/Writing	544	0.014	0.992	0.992
February 2026 Paper	Quantitative Reasoning	544	0.010	0.994	0.994

10.4 Validity Evidence Based on Item-Level Mode Effects

To further evaluate CLT's validity for the college-bound population, we used the same Mantel-Haenszel differential item functioning analyses in Chapter 9 to examine item-level mode effects for the February 2026 administration. Table 10.4 compares the in-school online and remotely-proctored modes, and Table 10.5 compares the in-school online and paper modes.

In the online versus remotely-proctored comparison, Verbal Reasoning showed the most mode-related DIF, with two B+ items, two B- items, and one C+ item. Grammar/Writing showed one B+ item and one B- item, and Quantitative Reasoning showed two B+ items and one B- item. In the online versus paper comparison, Verbal Reasoning showed one B+ item and three B- items, Grammar/Writing showed one B+ item, and Quantitative Reasoning showed no DIF in the 39 common items.

Only one C item was identified across the two mode comparisons, and the direction of DIF was mixed rather than consistently favoring one mode. This suggests that although some item-level mode effects were present in the college-bound sample, there is no evidence of a strong unidirectional effect that would be expected to distort score interpretation at the test level.

Table 10.4

DIF Results for In-School Online and Remotely-Proctored College-Bound Students

Form	Section	N_{Online}	N_{RP}	Total	A	B+	B-	C+	C-
February 2026	Verbal Reasoning	731	852	40	35	2	2	1	0
February 2026	Grammar/Writing	731	852	40	38	1	1	0	0
February 2026	Quantitative Reasoning	731	852	40	37	2	1	0	0

Note. The plus (+) sign means that the item favors remotely proctored students, the minus (-) sign means that the item favors in-school online students.

Table 10.5

DIF Results for In-School Online and Paper College-Bound Students

Form	Section	N_{Online}	N_{Paper}	Total	A	B+	B-	C+	C-
February 2026	Verbal Reasoning	731	544	40	36	1	3	0	0
February 2026	Grammar/Writing	731	544	40	39	1	0	0	0
February 2026	Quantitative Reasoning	731	544	39	39	0	0	0	0

Note. The plus (+) sign means that the item favors paper students, the minus (-) sign means that the item favors in-school online students. Quantitative Reasoning is based on 39 common items rather than 40 because one February 2026 paper item did not have an online counterpart and was excluded from the online-paper DIF analysis.

11. NORMING

This chapter provides national norms for CLT scores through the CLT-SAT® concordance study conducted in spring 2023 (Classic Learning Initiatives, 2023a). Specifically, we use the concordance table between the CLT and the SAT® to derive national percentile approximations for CLT scores based on the national SAT® percentiles (College Board, 2023). For more information on the SAT® national norms, the reader is referred to College Board (2023) and the SAT® Technical Manual (College Board, 2018).

In spring 2023, Classic Learning Initiatives conducted a concordance study which produced a new concordance table between CLT scores and SAT® scores. Given that the SAT® does not report separate scores for reading and writing, the Verbal Reasoning and the Grammar/Writing sections of the CLT were combined and linked to SAT® Evidence-Based Reading & Writing scores. Total CLT scores were linked to total SAT® scores, and Quantitative Reasoning scores were linked to SAT® Math scores. The sample that was used to link the total scores consisted of 4,375 students who took both the CLT and the SAT® in the past. The sample used to link the section scores consisted of 1,551 students. Equipercentile linking with a single group design was conducted to link the two scales. In this method, the linked scores have the same percentile rank within the same group of students. For two scales X and Y where X is linked to Y, this relationship is expressed in Equation 11.1 (Kolen & Brennan, 2014):

$$e_Y(x) = G^{-1}[F(x)] \tag{11.1}$$

where $e_Y(x)$ is the Y scale equivalent of score x, $F(x)$ is the cumulative distribution function of X, and G^{-1} is the inverse of the cumulative distribution function of Y.

Deriving national percentiles from the CLT-SAT® concordance assumes that the concordance link is strong. A strong linkage between two tests requires that: 1) the two tests measure similar constructs (Dorans & Walker, 2007); 2) both tests have high reliability (Dorans, 2004); and 3) the sample used to produce the concordance table is sufficiently large and representative of the target user population (Pommerich, 2007). Construct similarity is evaluated by analyzing the content of each test and by measuring the empirical relationship between the scores using correlations. Our concordance study demonstrated that the CLT and the SAT® measure highly similar constructs, as evidenced by: a) the results of our content alignment study which highlighted the similarities between the contents of the two tests – namely that both tests are measures of reading, writing, and mathematics; b) the high correlations between the scores (see Chapter 9 for the correlations) (Dorans, 2004; Dorans & Walker, 2007). Furthermore, Chapter 8 of this technical report showed that each section of the CLT is highly reliable, which is also true for each section of the SAT® (College Board, 2018).

Finally, we noted that the sample used to create the concordance tables must be both a) sufficiently large and b) representative of the intended population of users (Pommerich, 2007). A sufficiently large sample reduces the random error in the linked scores whereas a representative sample ensures that the concordance table does not systematically underestimate or overestimate the true concordance relationship. Kolen and Brennan (1995) suggested that a sample size of 1,500 provides sufficient precision for the equipercentile linking method, which our study exceeded. With respect to representation, our concordance study used data from the two groups of students who were the most likely to represent the future users of our concordance table: students who used the CLT in the past, and public school students whom we expect to comprise a larger proportion of our future users. To include public school students in the study, we organized a special CLT administration in March 2023 which provided 435 official CLT and SAT® scores from public school students.

In short, our concordance study fulfilled the criteria for a strong linkage between the CLT and the SAT®, showing that the two tests measure similar constructs, that both are highly reliable tests, and that the sample used to create the concordance table represented its intended users with a sufficient sample size. For these reasons, the concordance table was constructed with high confidence and can be leveraged to estimate national percentiles for the CLT. Table 11.1 shows each CLT total score, the corresponding SAT® total score, and the national SAT® percentile. Table 11.2 shows the CLT Verbal Reasoning + Grammar/Writing scores along with the corresponding SAT® scores and the SAT® national percentile. Table 11.3 shows the concordance between the CLT Quantitative Reasoning scores and SAT® Math scores with the corresponding SAT® national percentiles.

Table 11.1

The Concordance Between CLT and SAT Total Scores and the Corresponding SAT National Percentiles

CLT Total	SAT TOTAL	SAT NATIONAL PERCENTILE
120	1600	99+
119	1600	99+
118	1590	99+
117	1580	99+
116	1580	99+
115	1570	99+
114	1560	99+
113	1550	99+
112	1540	99+
111	1530	99+
110	1520	99+
109	1500	99
108	1490	99
107	1480	99
106	1470	99
105	1460	99
104	1440	98
103	1430	98
102	1420	98
101	1410	97
100	1390	97
99	1380	96
98	1370	96
97	1360	95
96	1340	94
95	1330	93
94	1320	93
93	1310	92
92	1300	91
91	1290	90
90	1270	88
89	1260	87
88	1250	86
87	1240	85
86	1230	84
85	1220	83
84	1210	82
83	1200	81
82	1190	80
81	1180	78
80	1170	77
79	1160	76
78	1150	74
77	1140	73
76	1140	73
75	1130	71
74	1120	70
73	1110	69
72	1100	67
71	1090	65
70	1080	63
69	1080	63
68	1070	61
67	1060	60
66	1050	58
65	1040	56
64	1040	56
63	1030	54
62	1020	52
61	1010	50
60	1000	48
59	1000	48
58	990	46
57	980	44
56	970	42
55	960	40
54	950	38
53	940	36
52	940	36
51	930	35
50	920	33
49	910	31
48	900	29
47	890	27
46	880	26
45	870	24
44	860	23
43	850	21
42	840	20
41	840	20
40	830	18
39	820	17
38	810	16
37	800	14
36	790	13
35	780	11
34	770	10
33	760	9
32	750	8
31	740	7
30	740	7
29	730	6
28	720	5
27	710	4
26	700	4
25	690	3
24	690	3
23	680	2
22	670	2
21	660	1
20	660	1
19	650	1
18	640	1
17	630	1
16	630	1
15	620	1-
14	610	1-
13	610	1-
12	600	1-
11	590	1-
10	590	1-
9	580	1-
8	570	1-
7	570	1-
6	560	1-
5	550	1-
4	550	1-
3	540	1-
2	530	1-
1	520	1-
0	510	1-

Table 11.2

The Concordance Between CLT Verbal Reasoning + Grammar/Writing and SAT EBRW Scores, and the Corresponding SAT National Percentiles

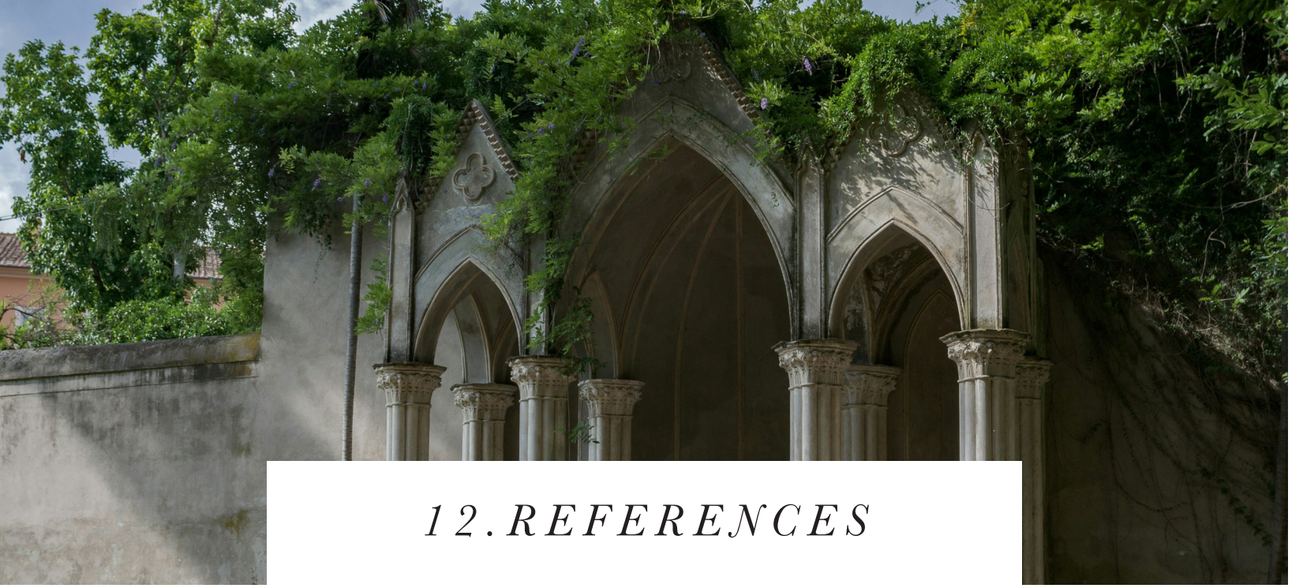
CLT VR + GW	SAT EBRW	SAT NATIONAL PERCENTILE
80	800	99+
79	790	99+
78	780	99+
77	770	99+
76	760	99+
75	750	99
74	740	99
73	730	99
72	730	99
71	720	98
70	710	97
69	700	97
68	690	96
67	690	96
66	680	95
65	670	93
64	670	93
63	660	92
62	650	90
61	640	88
60	640	88
59	630	86
58	620	84
57	620	84
56	610	81
55	600	79
54	600	79
53	590	76
52	580	74
51	580	74
50	570	71
49	560	68
48	560	68
47	550	65
46	540	62
45	540	62
44	530	58
43	520	55
42	520	55
41	510	51
40	510	51
39	500	48
38	490	44
37	490	44
36	480	41
35	470	38
34	470	38
33	460	34
32	450	31
31	450	31
30	440	28
29	440	28
28	430	24
27	420	22
26	420	22
25	410	19
24	400	16
23	400	16
22	390	13
21	380	11
20	380	11
19	370	9
18	360	7
17	360	7
16	350	5
15	340	3
14	340	3
13	330	2
12	320	2
11	320	2
10	310	1
9	300	1
8	290	1-
7	280	1-
6	280	1-
5	270	1-
4	260	1-
3	250	1-
2	230	1-
1	220	1-
0	210	1-

Table 11.3

The Concordance Between CLT Quantitative Reasoning and SAT Math Scores, and the Corresponding SAT National Percentiles

CLT Quantitative Reasoning	SAT MATH	SAT NATIONAL PERCENTILE
40	800	99+
39	790	99+
38	780	99
37	760	99
36	750	98
35	740	98
34	730	97
33	720	97
32	700	95
31	690	94
30	680	93
29	660	91
28	650	90
27	640	89
26	620	85
25	610	83
24	600	81
23	580	76
22	570	73
21	560	71
20	540	65
19	530	61
18	520	57
17	500	47
16	490	44
15	470	36
14	460	32
13	450	29
12	430	23
11	420	20
10	400	15
9	390	13
8	380	10
7	360	7
6	350	5
5	330	3

4	310	1
3	290	1-
2	270	1-
1	250	1-
0	220	1-



12. REFERENCES

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

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. BMJ, 314(7080), 572. https://doi.org/10.1136/bmj.314.7080.572

Cho, E., & Kim, S. (2014). Cronbach's coefficient alpha. Organizational Research Methods, 18(2), 207–230. https://doi.org/10.1177/1094428114555994

Classic Learning Initiatives. (2023a). The concordance relationship between the Classic Learning Test (CLT) and the Scholastic Aptitude Test (SAT). https://www.cltexam.com/wp-content/uploads/2023/04/2023-Concordance-Report.pdf

College Board. (2017). SAT Suite of Assessments Technical Manual. https://satsuite.collegeboard.org/media/pdf/sat-suite-assessments-technical-manual.pdf

College Board. (2023). Understanding SAT scores. https://satsuite.collegeboard.org/media/pdf/understanding-sat-scores.pdf

College Board. (2024). SAT® score relationships with college GPA: First-year through fourth-year cumulative GPA. https://research.collegeboard.org/media/pdf/SAT%28R%29%20Score%20Relationships%20with%20College%20GPA.pdf

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555

Custer, M., Sharairi, S., & Swift, D. (2012, April). A comparison of scoring options for omitted and not-reached items through the recovery of IRT parameters when utilizing the Rasch model and joint maximum likelihood estimation. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Dorans, N. J. (2004). Equating, concordance, and expectation. Applied Psychological Measurement, 28(4), 227–246. https://doi.org/10.1177/0146621604265031

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. ETS Research Report Series, 1992(1). https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), Linking and aligning scores and scales. Springer.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. Journal of Statistical Software, 48(4), 1–18. https://www.jstatsoft.org/article/view/v048i04

Florida Board of Governors. (2023). Top education system in the nation expands admission exam options with Classic Learning Test. https://www.flbog.edu/2023/09/08/top-education-system-in-the-nation-expands-admission-exam-options-with-classic-learning-test/

Florida Department of Education. (2023). Annual assessment requirement. https://www.fldoe.org/schools/school-choice/k-12-scholarship-programs/ftc/annual-assessment-requirement.stml

Florida Senate. (2023a). CS/CS/CS/HB 1537: Education. https://www.flsenate.gov/Session/Bill/2023/1537

Florida Senate. (2023b). 2023 Bill Summaries: CS/CS/CS/HB 1537 — Education. https://www.flsenate.gov/Committees/billsummaries/2023/html/1537

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. Educational and Psychological Measurement, 66(6), 930–944. https://doi.org/10.1177/0013164406288165

Harvill, L. M. (1991). Standard error of measurement. Educational Measurement: Issues and Practice, 10(2), 33–41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

Holland, P. W., & Thayer, D. T. (1985). An alternate definition of the ETS delta scale of item difficulty (ETS Program Statistics Research Technical Report No. 85-64). Educational Testing Service.

Kolen, M. J., & Brennan, R. L. (1995). Test equating: Methods and practices. Springer.

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking: Methods and practices. Springer.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices. Springer.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions, 16(2), 878.

Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. Educational and Psychological Measurement, 59(4), 615–630. https://doi.org/10.1177/0013164499594004

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. Psychological Methods, 1(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods, 42, 847–862. https://doi.org/10.3758/BRM.42.3.847

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4), 719–748.

Mislevy, R. J., & Wu, P.-K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing (ETS Research Report No. RR-96-30-ONR). ETS Research Report Series, 1996(2). https://doi.org/10.1002/j.2333-8504.1996.tb01708.x

Pommerich, M. (2007). Concordance: The good, the bad, and the ugly. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), Linking and aligning scores and scales. Springer.

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2025). psych: Procedures for psychological, psychometric, and personality research (Version 2.5.3) [Computer software]. Northwestern University. https://CRAN.R-project.org/package=-psych

Robitzsch, A., Kiefer, T., & Wu, M. (2024). TAM: Test Analysis Modules (Version 4.3-4) [Computer software]. https://CRAN.R-project.org/package=TAM

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. Psychometrika, 82(3), 795–819. https://doi.org/10.1007/s11336-016-9544-7

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Shin, A.-Y. (2009). Investigating the effects of missing data treatments on item response theory vertical scaling [Doctoral dissertation, University of Iowa].

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education, 48, 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Texas Education Agency. (2022). Standard technical processes. In 2021–2022 Technical Digest (Chapter 3). https://tea.texas.gov/student-assessment/testing/2021-2022-technical-digest-chapter-3.pdf

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38(1), 1–10. https://doi.org/10.1007/BF02291170

von Davier, M., & von Davier, A. A. (2004). A unified approach to IRT scale linking and scale transformations (ETS Research Report No. RR-04-09). ETS Research Report Series, 2004(1). https://doi.org/10.1002/j.2333-8504.2004.tb01936.x

Welton, G. L. (2025). Classic Learning Test (CLT) as a predictor of student performance. Grove City College. https://www.gcc.edu/Portals/0/CLT-Report_1025.pdf

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. MESA Press.

Zieky, M. (2003). A DIF primer. Educational Testing Service. https://www.ets.org/content/dam/ets-org/pdfs/praxis/dif-primer.pdf

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. ETS Research Report Series, 2012(1). https://doi.org/10.1002/j.2333-8504.2012.tb02290.x