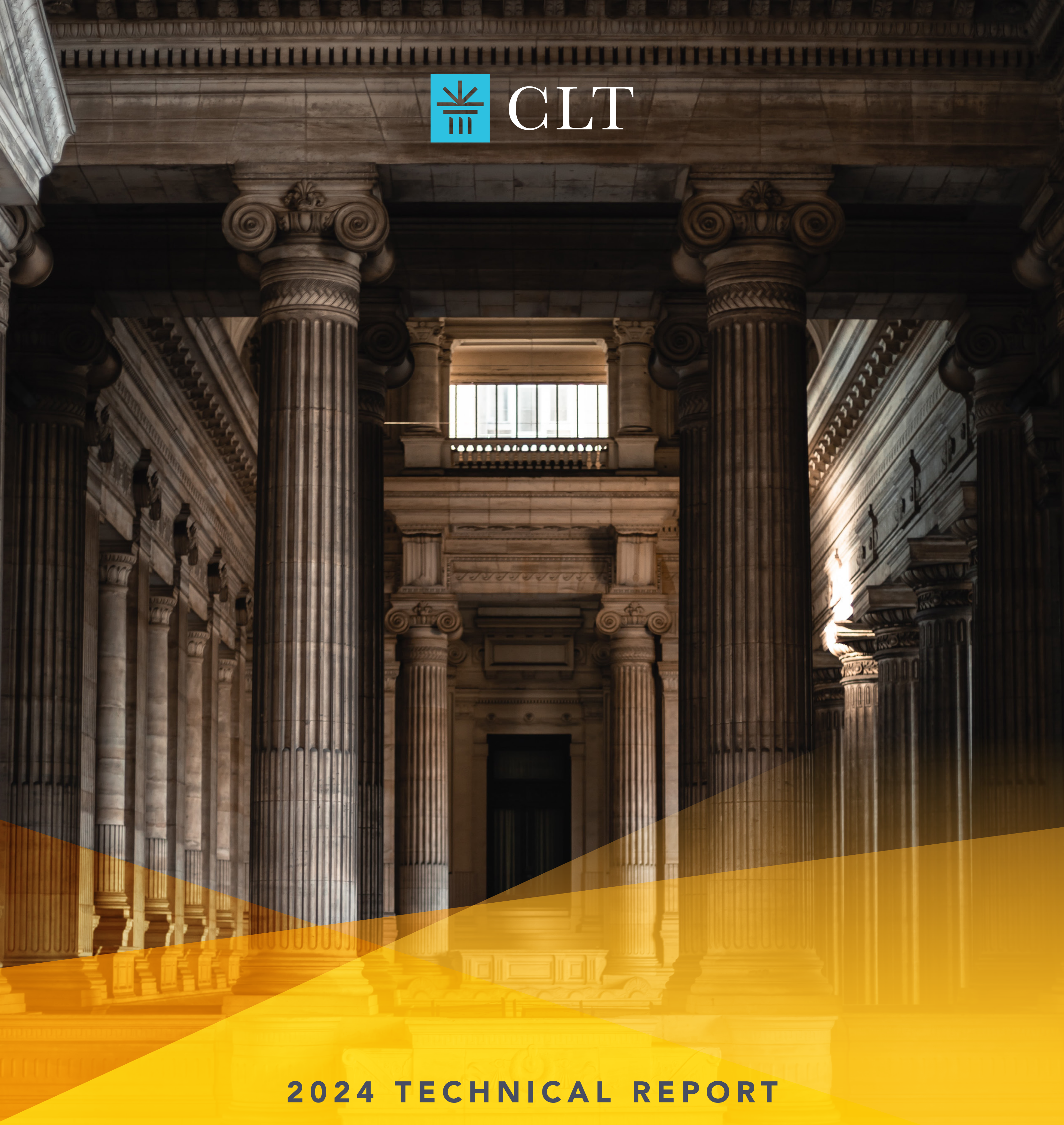


CLT

2024 TECHNICAL REPORT

The Classic Learning Test

CLT

2024 TECHNICAL REPORT

The Classic Learning Test

CONTRIBUTORS

Report Lead .... Noah Tyler

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

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

Dr. Todd Johnson
Associate Professor, Department of Leadership and Teacher Education, University of South Alabama; Ph.D., Special Education and Research Methods, George Mason University

Writers & Reviewers ...... Livvy Beaver
Natalie Walker
Mary Trent
Gabriel Blanchard

Graphic Designer .... Meg Pilcher

Letter from the CEO

Could there be anything more boring than standardized testing? I think the influence of standardized tests has flown under the radar because of just how boring these high-stakes tests are, and how inconsequential their content and questions feel, when in fact they are levers of curriculum change.

By the time I entered school in the mid 1980s, any question that carried moral or ethical implications, or any question about the purpose of life, sacred responsibilities, or where to find human happiness, had been removed from the classroom. The education I experienced had been designed with purely utilitarian ends in mind. Any transcendent idea had been gutted from the curriculum and as a result, like most of my classmates, I was painfully bored.

It wasn't until graduate school that I came to appreciate the holistic education previous Americans had received. The founding fathers of the United States revived my imagination. They were deeply interested in philosophy, human nature, political theory, and the pursuit of happiness. The education they received was aimed, most fundamentally, at making a person more fully human.

As I questioned how such a beautiful concept of education had been lost, I came to the conclusion that high-stakes testing was partially to blame. Shouldn't the most important tests that students take also engage those students with some of the most important ideas, texts, and subjects? CLT was born in response to this question with three core values:

Anchored, because we hope that we can be a catalyst for renewal in education nationwide by offering a new standard that puts students in front of the thinkers and questions that have most meaningfully shaped our culture for the past two millennia. After all, education is always a renewal: one where cumulative, cultural memory is awakened in a student's personal memory, where it lives a life of its own. E. D. Hirsch has called it "acculturation"; Eva Brann has called it "a privately performed renaissance."

Passionate, because that's the only form such a renewal– where the survival and richness of both the inner lives of individuals and the future of our culture are at stake– can take.

Humane, because we believe the fundamental goal of education is to make a person more fully human. Thus, we seek to make every CLT interaction humanizing, and that is also our hope for your experience with this Technical Report.

Jeremy Tate

Jeremy Tate,
Founder and CEO of CLT

TABLE OF CONTENTS

Chapter 1: Introduction 1

1.1 What is the CLT?

1.2 Improving Students' Test-Taking Experience

1.3 Motivating Positive Change in Assessment and Education

1.4 CLT in Context

1.5 About the CLT Technical Report

Chapter 2: Standards and Content Coverage 7

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

2.2 Author Bank

2.3 Verbal Reasoning Test

2.4 Grammar/Writing Test

2.5 Quantitative Reasoning Test

2.6 Optional Essay

Chapter 3: Test Development 23

3.1 Test Blueprint

3.2 Selecting and Training Item Developers

3.3 Test Form Creation

3.4 Quality Control Procedures

3.5 Licensing and Permissions

Chapter 4: Test Administration 29

4.1 Overview

4.2 Test Modes

4.3 Test Day Processes and Procedures

4.4 Test Day Schedules

4.5 Test Day CLT Support

4.6 Test Security

Chapter 5: Test Accessibility 39

5.1 Fairness During the Testing Process

5.2 Fairness in Test Accessibility

5.3 Accommodations and Requests

Chapter 6: Test Results 43

6.1 Student Score Reports

6.2 College Score Reports

6.3 Secondary School Score Reports

Chapter 7: Classical Item Analysis...................................... 49

7.1 Introduction to Item Statistics in Classical Test Theory

7.2 The Results of Classical Item Analysis

Chapter 8: IRT Calibration, Scaling, and Scoring 53

8.1 Introduction

8.2 The Rasch Model

8.3 Fixed Parameter Calibration Using the Common Item Nonequivalent Groups Design

8.4 Scaling and Reported Scores

Chapter 9: Reliability.. 69

9.1 Introduction

9.2 Quantifying Reliability

9.3 The Reliability and SEM of the Base Form

9.4 Test Information and Conditional Standard Errors of Measurement

Chapter 10: Validity.. 79

10.1 What is Validity?

10.2 Sources of Validity Evidence

10.3 Validity Evidence Based on Internal Structure: Confirmatory Factor Analysis

10.4 Validity Evidence Based on Internal Structure: Differential Item Functioning

10.5 Validity Based on Convergent Evidence: The Relationship Between the CLT and the SAT

Chapter 11: Norming 93

References.. 99

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 one hundred thousand CLT assessments have been administered in homes and schools across the United States,[1] and over two hundred colleges and universities have adopted it as an admissions test.[2]

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."

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.

> **"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.

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.

Each section loads into a single browser window, so students can scroll to any part of that section without changing pages. A progress bar is provided at the top of the page, giving students a visual sense of their progress on the exam.

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.)

SHORTER TEST—FASTER 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 following Wednesday. 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 on the third Wednesday following 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

CHAPTER TWO

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 principal 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 for college readiness.

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 orderly 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.

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

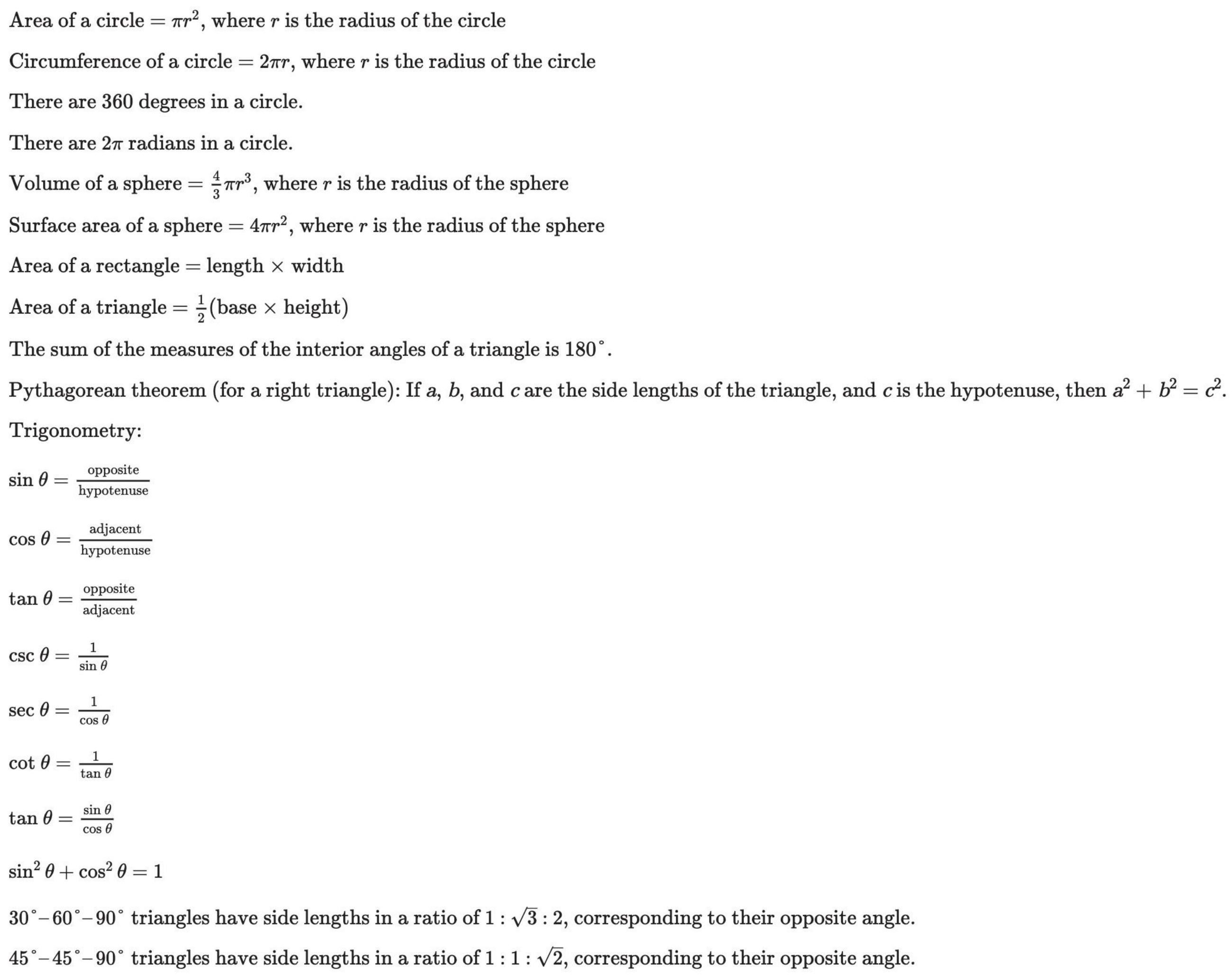
Passage adapted from St. Teresa of Ávila's The Way of Perfection, *1583.*

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, both at the top of the section and accessible at any time by selecting the *f(x)* 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 × width

Area of a triangle $= \frac{1}{2}$(base × 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?

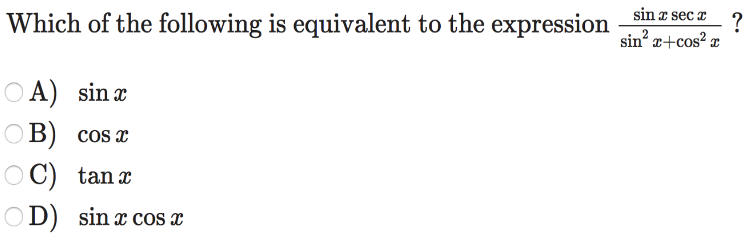
○ A) $y = \frac{3}{2}x - 7$
○ B) $y = -\frac{1}{6}x + 7$
○ C) $y = -\frac{2}{3}x + 7$
○ D) $y = \frac{3}{2}x - \frac{1}{7}$

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 $\frac{\sin x \sec x}{\sin^2 x+\cos^2 x}$?

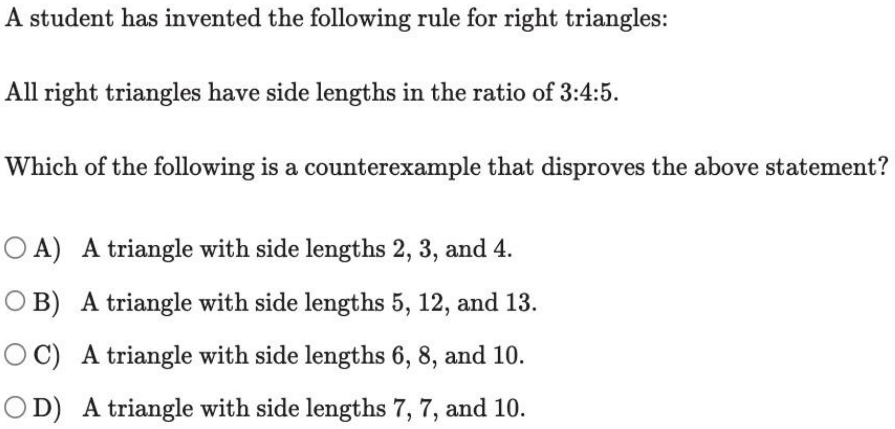
○ A) $\sin x$

○ 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?

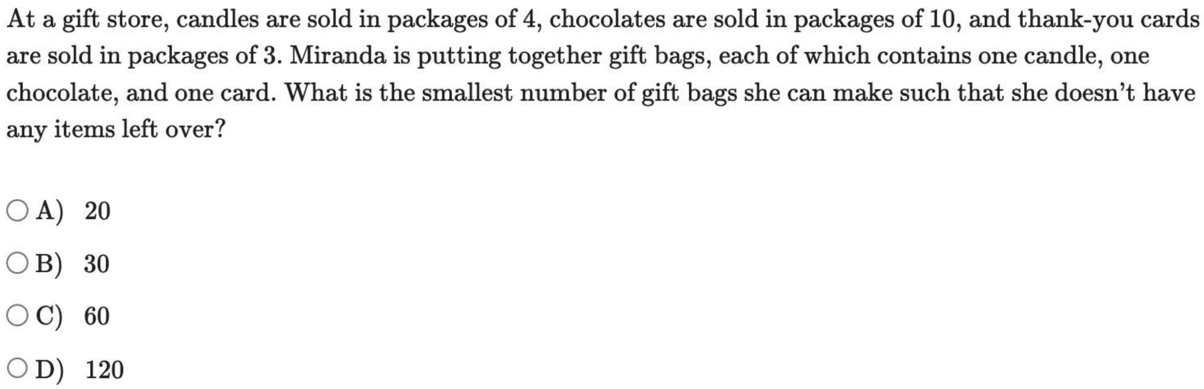
○ A) A triangle with side lengths 2, 3, and 4.

○ 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 Test Development team develops new test forms for every test date in conformity with our test blueprints, described in the previous chapter of this report, and statistical parameters, described in Chapter 8 of this report.

When developing forms that fit the blueprint, the team considers criteria such as content domain, complexity, and accessibility to the population of test-takers. The team's psychometrician(s) weigh in on whether each passage and item within the blueprint meet the desired psychometric properties defined for the CLT.

Many items have already undergone multiple rounds of review by expert judges in past years. Many test forms that follow the test blueprint have already been shown to perform well. These forms are available for reuse, assuming licensing is up to date and no errors have been found in any items. Test forms are studied post-test to guarantee that blueprint content specifications were met and that mastery of the targeted content was reflected in test scores.

3.2 Selecting and Training Item Developers

The CLT Test Development team chooses an item writer based on his or her qualifications and demonstrated ability in particular subject areas; many have grade-level experience in fields such as teaching and tutoring. New item writers are supervised by experienced members of the Test Writing team and are trained on the breakdown of question types, difficulty levels, and house style of the CLT suite of exams. Their work then goes through multiple rounds of revision and editing to ensure that each section maintains the high standards of the CLT, and that it is consistent, clear, and accurate.

CLT editorial reviewers have strong content knowledge in the areas of reading comprehension, grammar and writing, and/or math and logic, in addition to a keen eye for finding mistakes, typographical errors, inconsistencies, or stylistic issues. Occasionally, expert reviewers are also asked to review item scoring post-test to ensure items are appropriate.

3.3 Test Form Creation

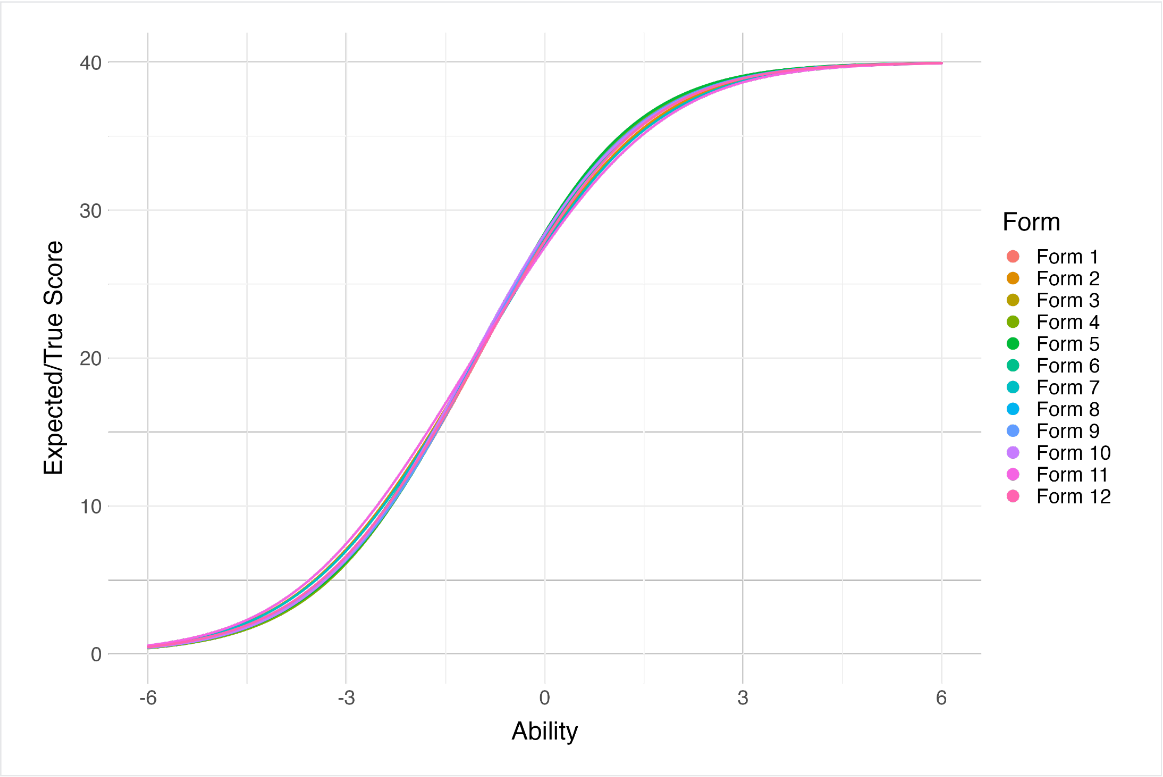
TEST FORM ASSEMBLY

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 and total scores, as explained in Chapter 7. 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.

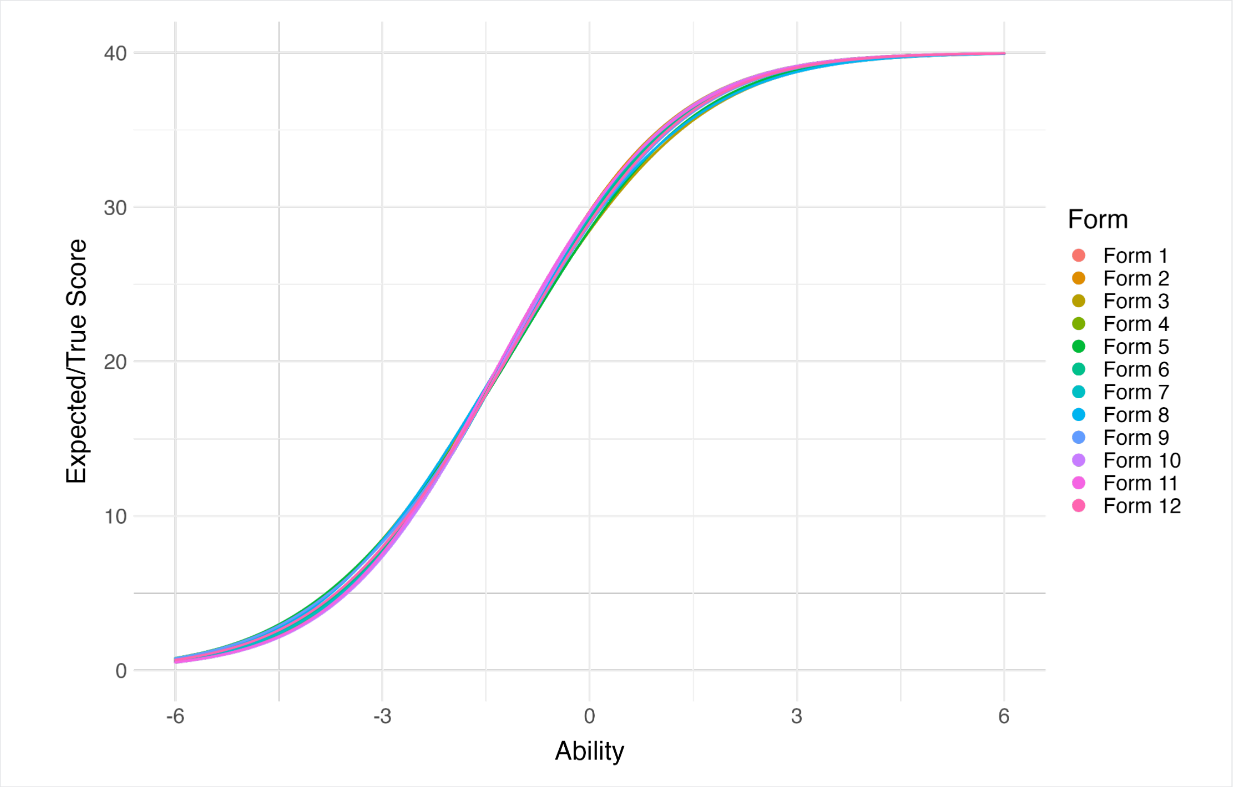
In addition to ensuring that only high-quality items are used, ATA allows us to construct forms that have a consistent level of difficulty. This is accomplished by defining an *objective function*, which is the statistical outcome that the ATA algorithm strives to achieve. For example, test information 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 12 modules assembled for the 2022-2023 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. A high overlap between the curves means that the difficulty differences between the modules are small. 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

The Test Characteristic Curves of the 2022-2023 Forms

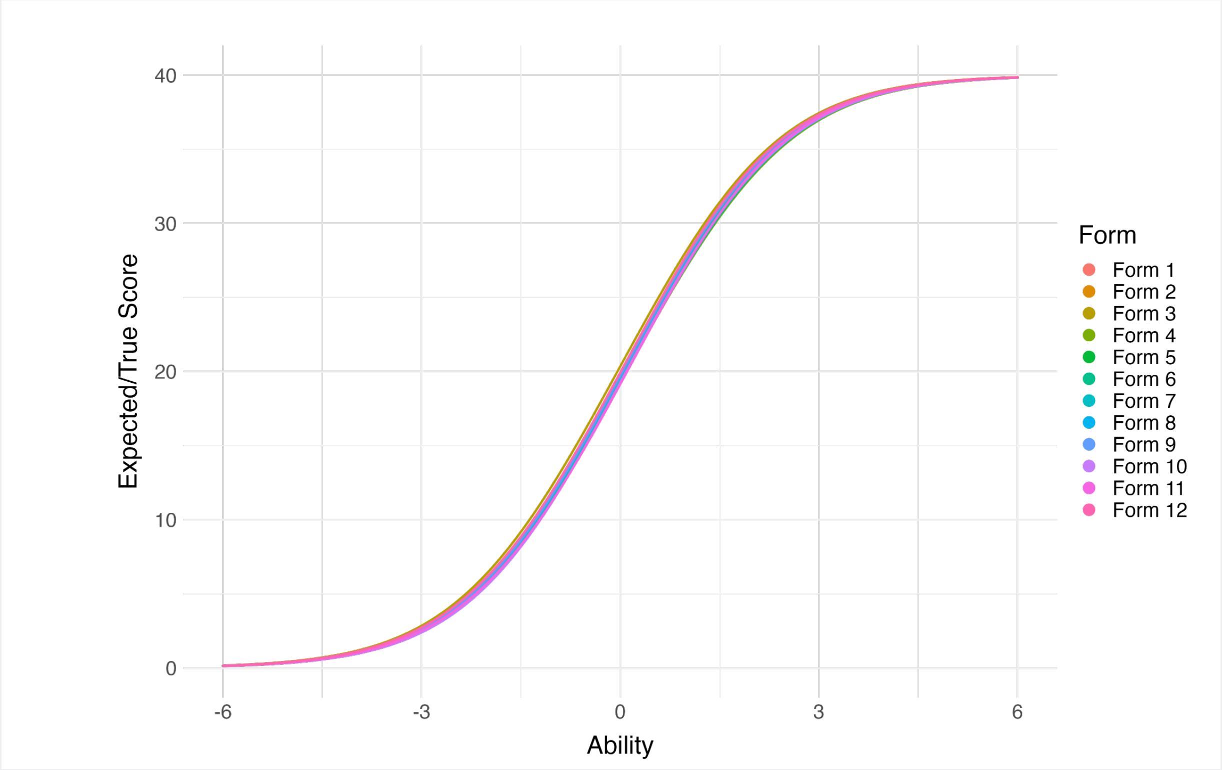


a) The test characteristic curves of the 2022-2023 Verbal Reasoning modules. The x-axis shows a range of ability levels on the logit scale. The y-axis shows the expected score of students with a given ability level. Each line represents the expected scores on a single form.



b) The test characteristic curves of the 2022-2023 Grammar/Writing 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.

CHAPTER THREE



c) The test characteristic curves of the 2022-2023 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.

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. Starting in the 2024-2025 academic year, large print versions will be 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.

ITEM BANK CALIBRATION

At the end of each academic year, the item bank is recalibrated and the item difficulties, discrimination values, and the model fit indices are updated. This newly calibrated bank of items are used to construct forms for the next academic year. Details of IRT calibration and item analyses are provided in Chapter 8.

3.4 Quality Control Procedures

ITEM BANK

CLT maintains a running item bank in order to track individual item use. The item bank is thoroughly evaluated by an internal reviewer to ensure that no inactive items are used on test forms, items include the correct supplementary information, and no errors are present. The CLT item banks are secure and only accessible to employees with privileged access, ensuring that no active items are made available to students outside of test day. With the creation of each test form, items that functioned differently on past forms are flagged. These items then undergo multiple review processes in which reviewers consider item data and actual vs. intended difficulty level. Reviewers either advocate for editing items or replacing them. Items that are replaced are made inactive within the item bank so that they cannot be reused. This item flagging process ensures that all active items within the item bank follow CLT's item quality standards.

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, blueprint checks are provided as an output of the automated test assembly algorithm and are 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.

Finalization procedures include an answer key check, a check of all supplementary question information, a check that all intended edits were made, and final approval from the Test Development team. After finalization, forms are passed to Operations for administration.

3.5 Licensing and Permissions

The Rights and Permissions team secures rights for any passages or images at least eight weeks in advance of each exam. The team works with publishers, agents, and authors to secure rights for all passages or images under copyright. Licenses are secured for usage across multiple years and administrations. CLT typically requests worldwide rights to accommodate international test takers. Longer term contracts are sought (three years or more) to maximize reuse from the item bank. Licensing requests include the right to make minor punctuation changes to fit testing context, minor spelling changes to modernize certain words, and the addition of line numbers to guide students in the test. Licensing rights and permissions are tracked and maintained in a secure database. Licenses are renewed when deadlines or print and online copy limits are reached.

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 booklet. 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 and browsers. It requires a reliable internet connection with Javascript enabled. Questions in the Quantitative Reasoning portion of the exam may include mathematical notation. Mathematical notation is scripted in HTML (MathML), and is visible regardless of the browser used to take the exam.

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 Wednesday following 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.

Password - In order to take the exam on test day, students must enter the proctor password specific to the exam in question. Proctors receive the password directly from CLT the week before and the day before test day. They provide their students with this password 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 - One of the proctors’ primary duties is to ensure that all students adhere to the designated time lengths for each of the exam’s sections. 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 CLT Administration 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 or open an additional page
- 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. ***Enforcing Section Times:*** The proctor is responsible for keeping time for each section. All sections of the exam must be completed within the allotted times. The proctor cannot lengthen the standard times for any of the test's sections unless the student has received testing accommodations approval from CLT.
4. ***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.
5. ***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.
6. ***Completing Administration Report and Proctor Survey:*** Immediately after the exam, proctors should fill out and submit an Administration Report and Proctor Survey which certifies their adherence to the rules and procedures of the exam administration and 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.

The Test Access Code 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.
   - Chrome or Firefox are the only supported browsers.
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 timing 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 (INSCHOOL 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.
10:55/10:59 AM	Proctor gives 5 minutes/1 minute warnings for Section 1.
11:00 AM	End of Verbal Reasoning section, beginning of Grammar/Writing section.
11:30/11:34 AM	Proctor gives 5 minutes/1 minute warnings for Section 2.
11:35 AM	End of Grammar/Writing section, beginning of restroom break.
11:45 AM	End of restroom break, beginning of Quantitative Reasoning section.
12:25/12:29 PM	Proctor gives 5 minutes/1 minute warnings for Section 3.
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.
1:00/1:04 PM	Proctor gives 5 minutes/1 minute warnings for the essay.
1:05 PM	End of the optional essay, dismissal of remaining students.
1:10 PM	Proctor submits Administration 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, teachers, and parents. 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

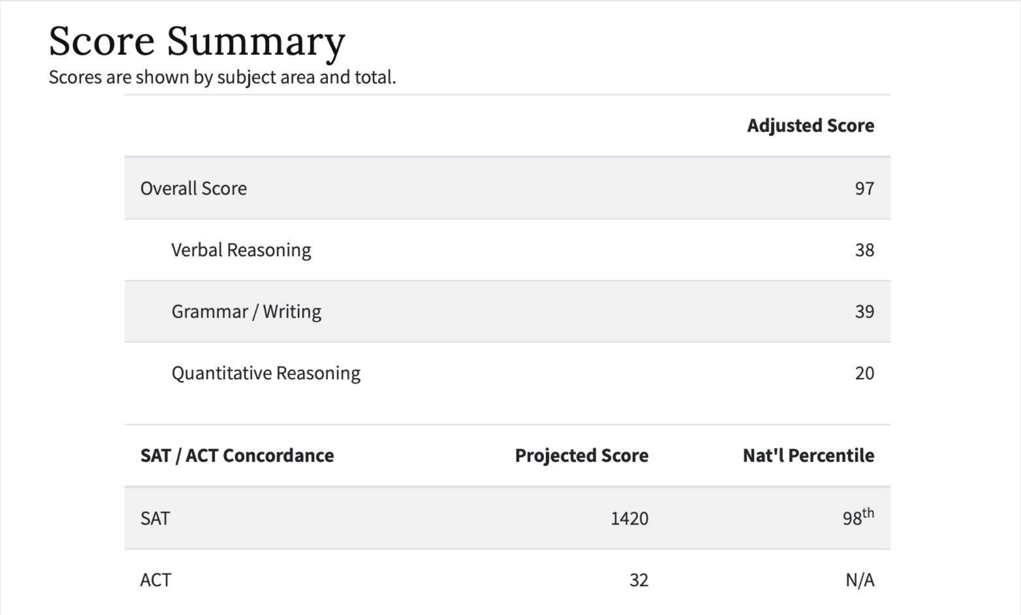
6.1 Student Score Reports

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

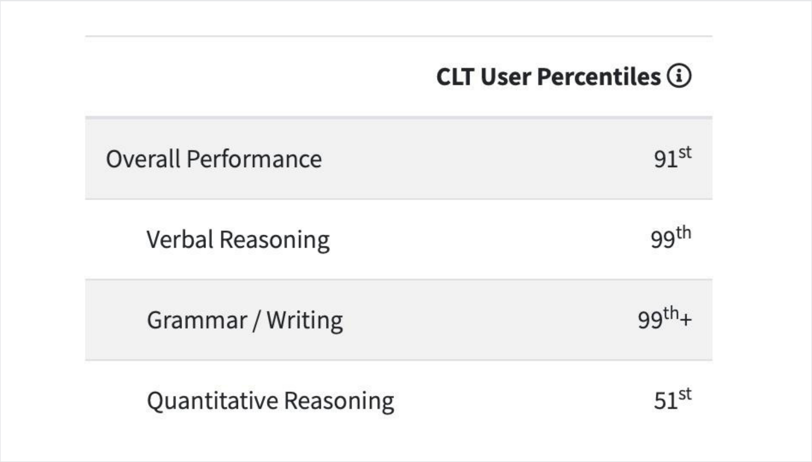
Scores are shown by subject area and total.

	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

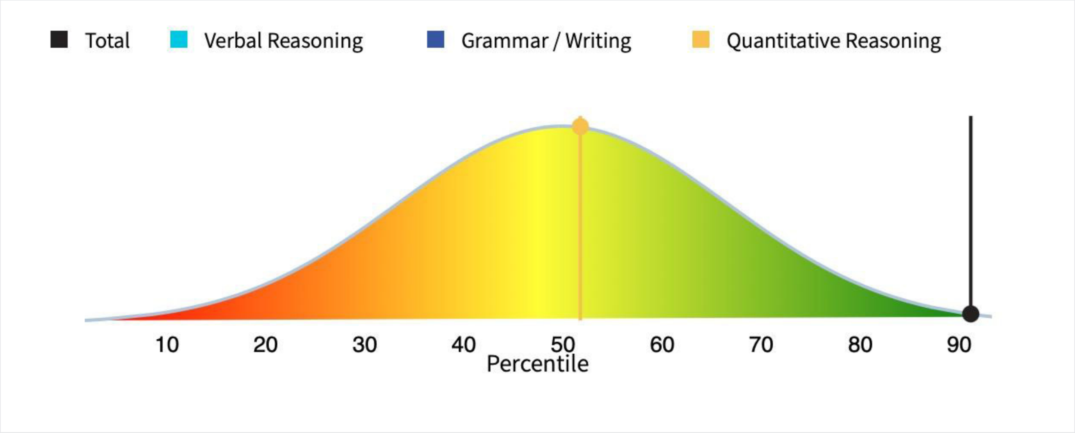
CLT User Percentiles show the percentage of CLT scores that are equal to or below the tester's score.



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

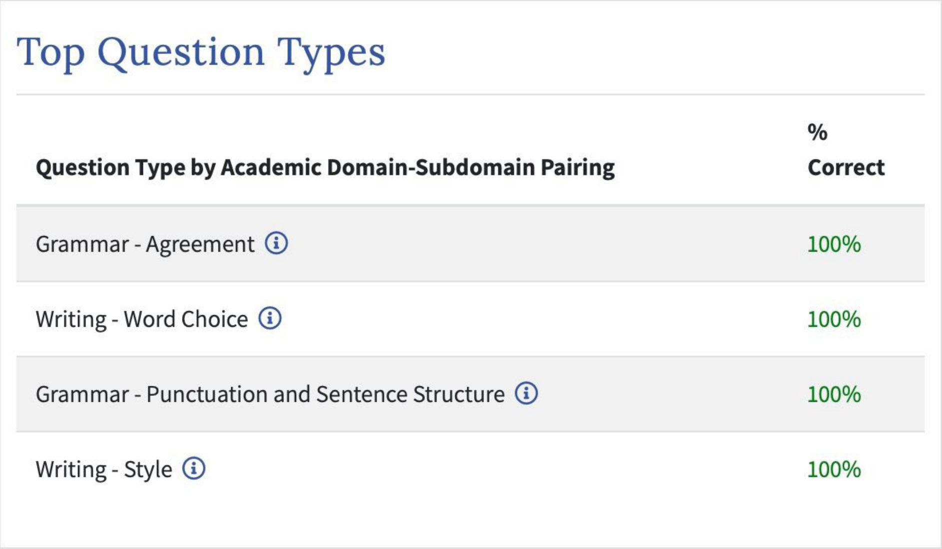
3. CLT SCORE BELL CURVE

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

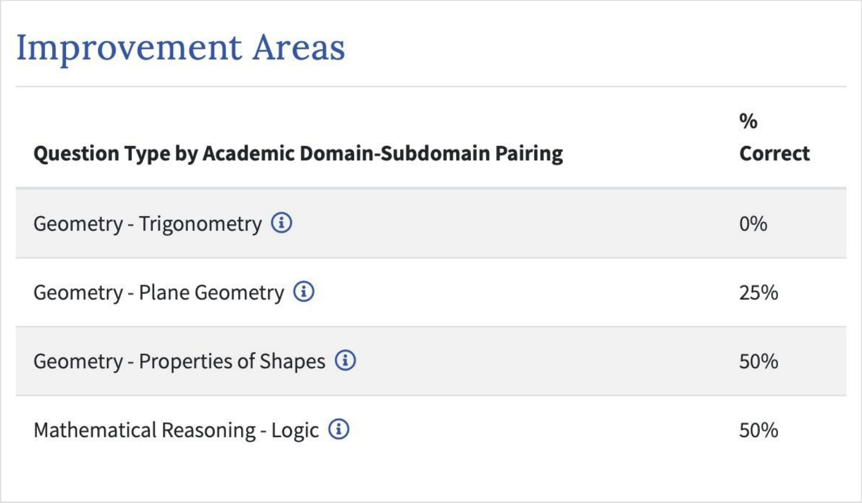
The Domains and Subdomains Report shows tester strengths and potential weaknesses. "Top Question Types" shows the types of questions which were answered with the highest accuracy.



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.



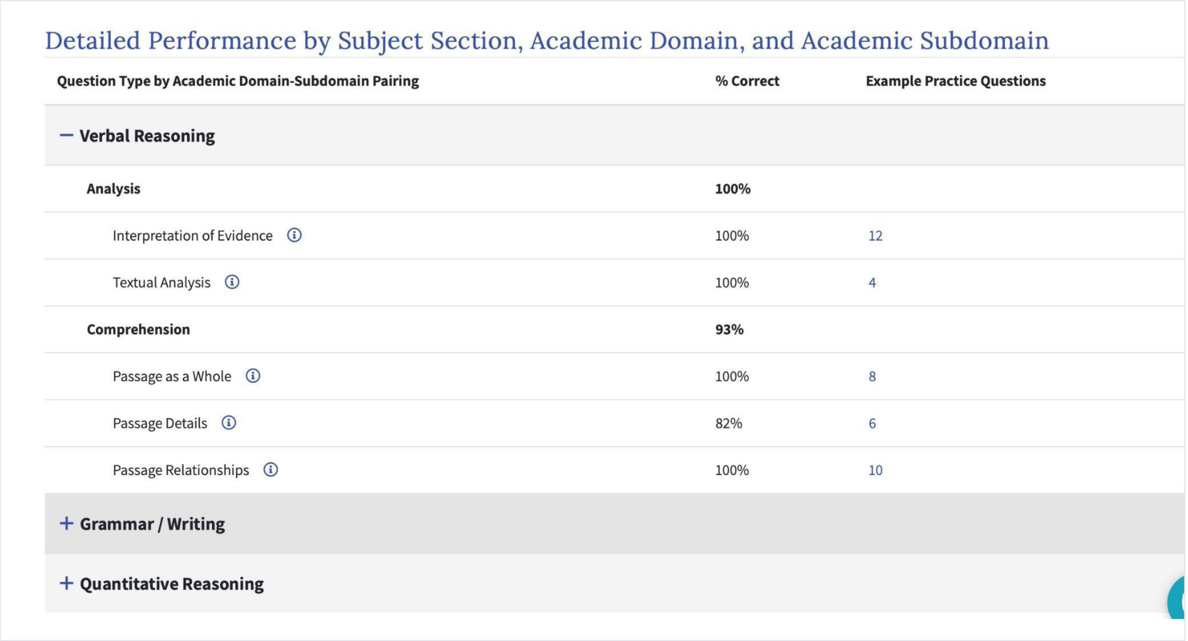
Improvement Areas

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.



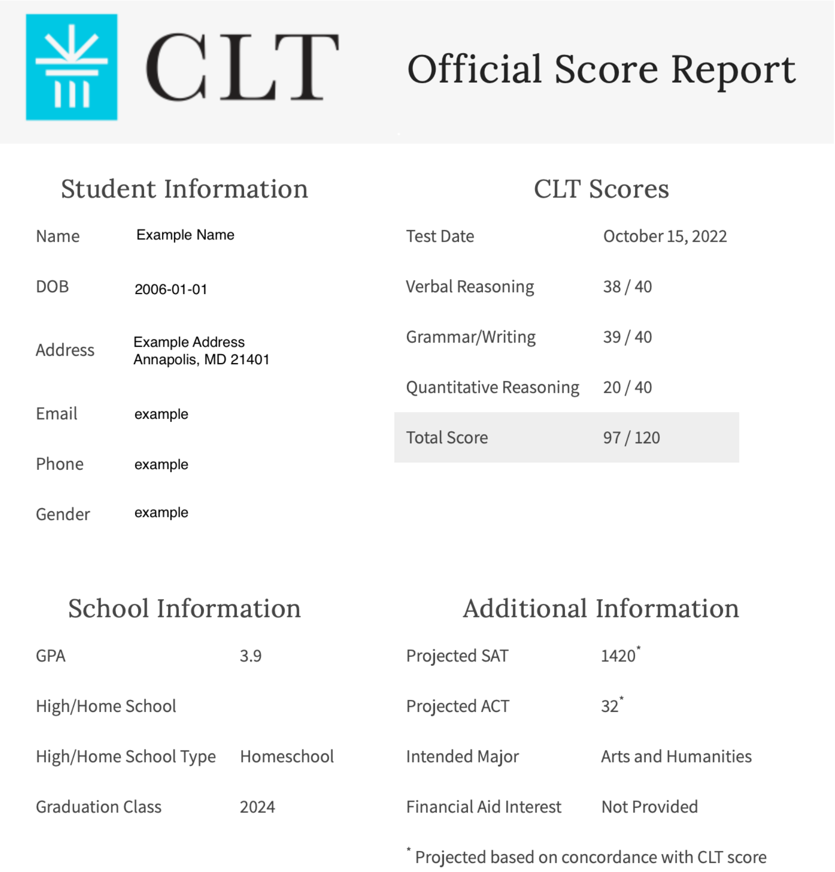
Detailed Performance by Subject Section, Academic Domain, and Academic Subdomain

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

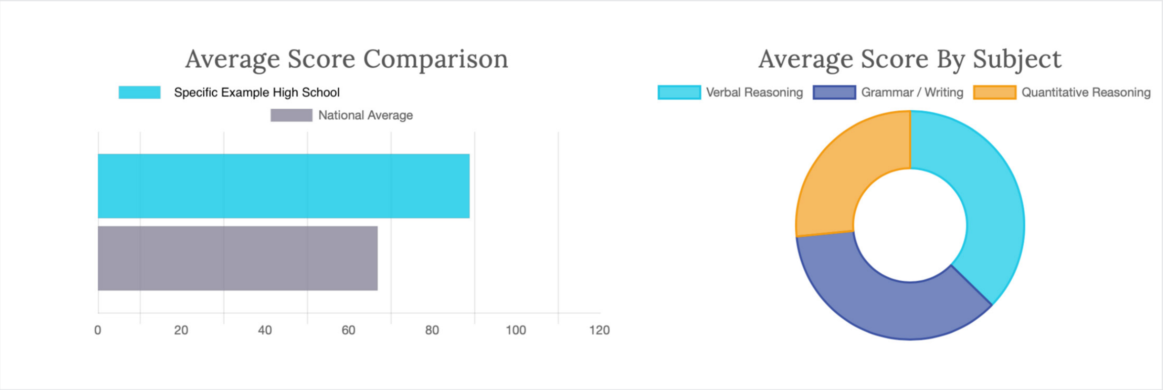
* Projected based on concordance with CLT score

6.3 Secondary School Score Reports

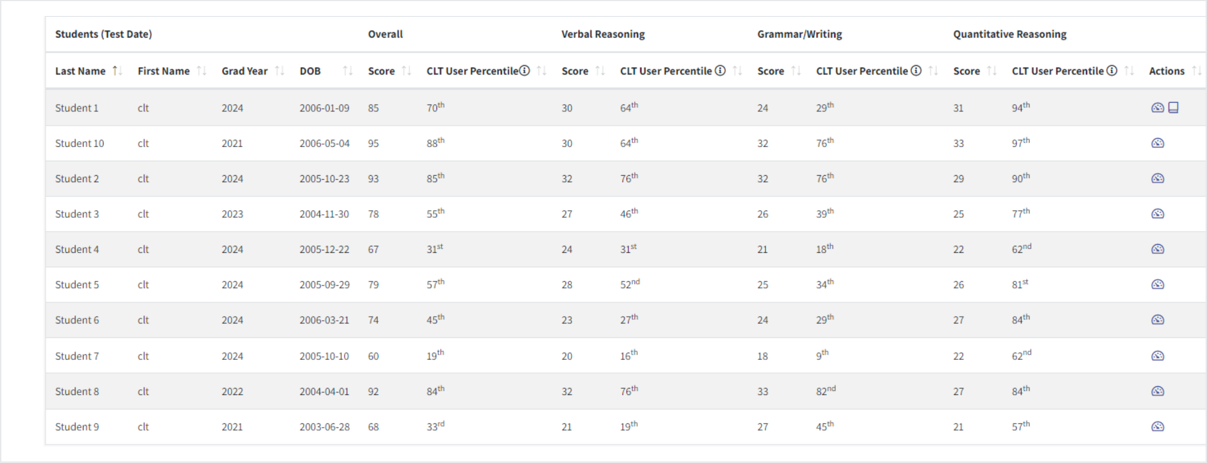
CLT provides detailed class and individual analytics for schools who have offered the CLT exam.

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	

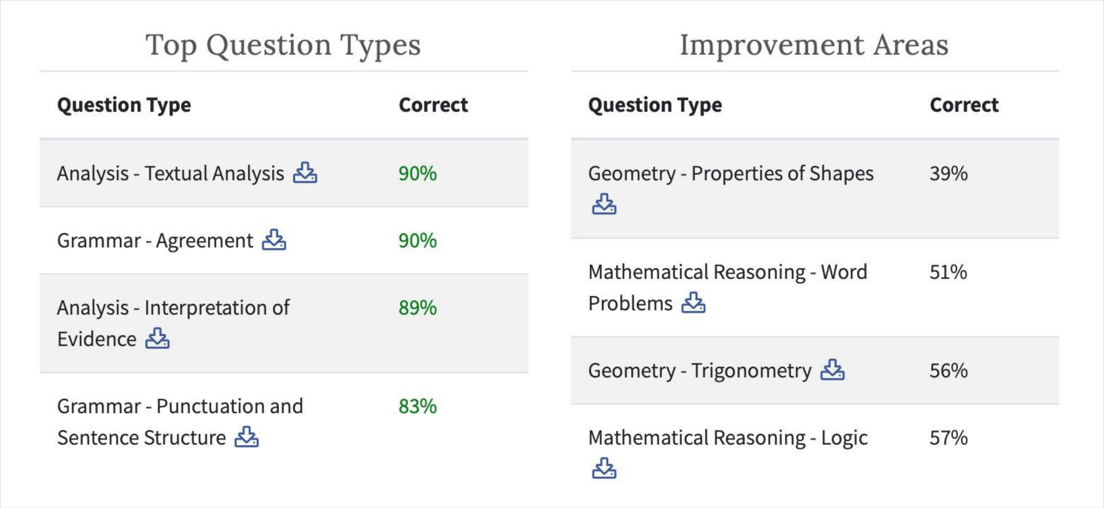
Test administrators also have access to detailed student and school-level analytics. Student performance is reported individually by academic domain and academic subdomain.

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.

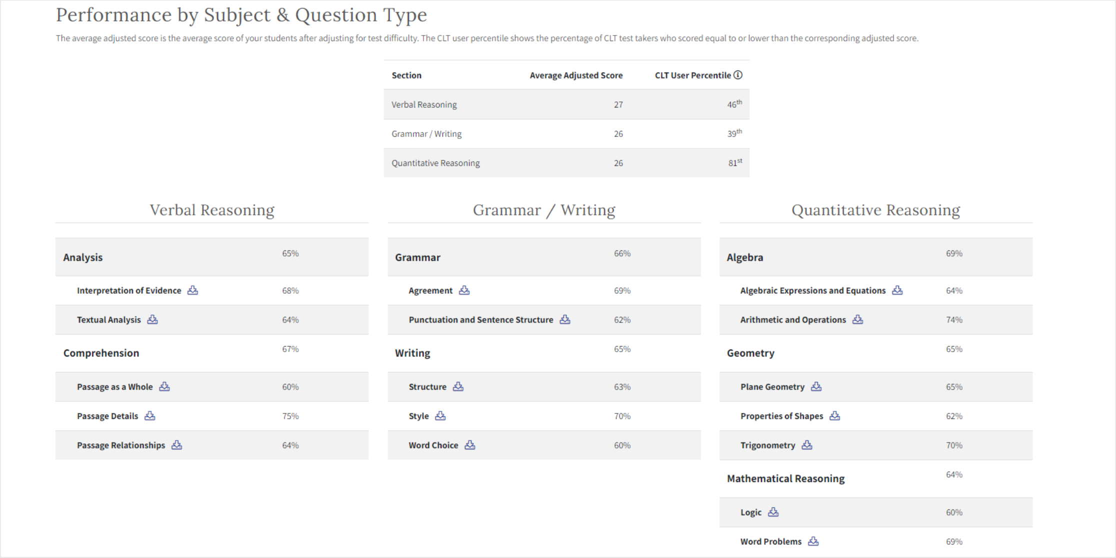


Top Question Types

Question Type	Correct
Analysis - Textual Analysis	90%
Grammar - Agreement	90%
Analysis - Interpretation of Evidence	89%
Grammar - Punctuation and Sentence Structure	83%

Improvement Areas

Question Type	Correct
Geometry - Properties of Shapes	39%
Mathematical Reasoning - Word Problems	51%
Geometry - Trigonometry	56%
Mathematical Reasoning - Logic	57%



Performance by Subject & Question Type

The average adjusted score is the average score of your students after adjusting for test difficulty. The CLT user percentile shows the percentage of CLT test takers who scored equal to or lower than the corresponding adjusted score.

Section	Average Adjusted Score	CLT User Percentile ⓘ
Verbal Reasoning	27	46th
Grammar / Writing	26	39th
Quantitative Reasoning	26	81st

Verbal Reasoning

Analysis	65%
Interpretation of Evidence	68%
Textual Analysis	64%
Comprehension	67%
Passage as a Whole	60%
Passage Details	75%
Passage Relationships	64%

Grammar / Writing

Grammar	66%
Agreement	69%
Punctuation and Sentence Structure	62%
Writing	65%
Structure	63%
Style	70%
Word Choice	60%

Quantitative Reasoning

Algebra	69%
Algebraic Expressions and Equations	64%
Arithmetic and Operations	74%
Geometry	65%
Plane Geometry	65%
Properties of Shapes	62%
Trigonometry	70%
Mathematical Reasoning	64%
Logic	60%
Word Problems	69%

7. CLASSICAL ITEM ANALYSIS

CHAPTER SEVEN

7.1 Introduction to Item Statistics in Classical Test Theory

This chapter introduces psychometric measures used to analyze test items in the Classical Test Theory (CTT) framework, specifically item difficulty and item discrimination. We show the results of CTT item analysis for the base CLT form that is also used as a reference form in the other psychometric analyses presented in this technical report.

In CTT, the difficulty of an item is defined as the proportion of test takers who answered the item correctly, also referred to as the p value (Crocker & Algina, 2008). The p value of item i is given by Equation 7.1:

$$p_i = \frac{\sum\limits_{n=1}^{N} X_{ni}}{N} \tag{7.1}$$

where X_{ni} is the response of student n to item i, coded as 1 for a correct answer and 0 for an incorrect answer. N is the total number of students. The p value is sample dependent, meaning it will change depending on the abilities of a sample of test takers; the same item may appear as difficult in a low-ability sample but easy in a high-ability sample (Chapter 8 presents difficulty metrics that are comparable across samples).

Item discrimination is the ability of an item to distinguish students of high ability from those of low ability (Crocker & Algina, 2008). In CTT, one measure of item discrimination is the point-biserial correlation between the responses to an item and the total scores on the test (the calculation of the total scores excludes the item being analyzed). The point-biserial correlation (r_{pb}) is the correlation between a binary variable (item responses) and a continuous variable (total scores), ranging 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. Thus, the correlation between 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. Like the p value, the point-biserial correlation is sample dependent, meaning that the point-biserial correlation of an item may vary across groups. The p values and point-biserial correlations are calculated using the CTT package (Willse, 2022) in the R programming language (R Core Team, 2023).

7.2 The Results of Classical Item Analysis

Table 7.1. shows the distribution of p values in the base form, and Table 7.2. shows the distribution of the point-biserial correlations.

Table 7.1

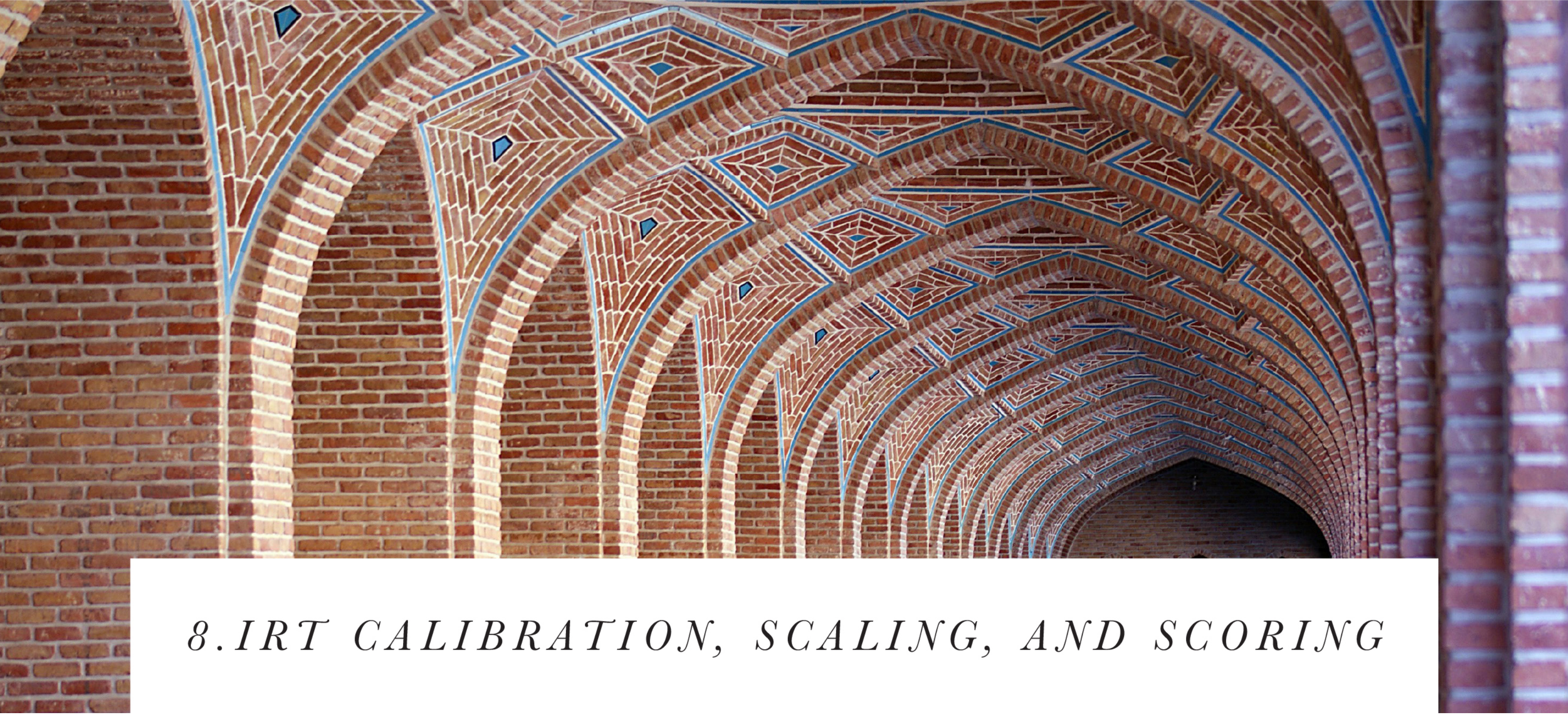
The Distribution of p Values in the CLT Base Form

Section	N Items	Mean	SD	Min	Max
Verbal Reasoning	40	0.64	0.17	0.14	0.96
Grammar/Writing	40	0.65	0.19	0.28	0.94
Quantitative Reasoning	40	0.49	0.17	0.18	0.82

Table 7.2

The Distribution of Point-Biserial Correlation r_{pb} in the CLT Base Form

Section	N Items	Mean	SD	Min	Max
Verbal Reasoning	40	0.35	0.09	0.16	0.50
Grammar/Writing	40	0.34	0.10	0.12	0.52
Quantitative Reasoning	40	0.35	0.09	0.13	0.51



8. IRT CALIBRATION, SCALING, AND SCORING

8.1 Introduction

CLT develops multiple test forms using different 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 every test-taker is scored fairly and consistently. For example, two students who took different forms but have the same ability in terms of the latent construct that the test measures 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 achieve these objectives, CLT conducts a series of psychometric analyses based on the modern measurement theory, Item Response Theory (IRT). IRT consists of a family of latent variable models that model the probability of a correct response to an item based on the interaction between item parameters and latent ability parameters. Item and ability parameters obtained from different test forms are placed on the same scale through a process of calibration that is described below (Kolen & Brennan, 2014). Using IRT enables two key outcomes: 1) measurement of an individual student's ability that is independent of the items on a particular test form, and 2) evaluation of test items that is independent of any particular group of test-takers. However, the scale of latent ability estimates obtained from IRT models is hard to interpret, so the ability estimates are transformed to scale scores that can be interpreted and understood more easily by stakeholders.

This chapter begins with an overview of the IRT model CLT uses, called the Rasch model. Then, we explain the calibration process that is carried out to ensure that the ability estimates obtained from the Rasch model can be compared across test forms and groups of students. 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.

8.2 The Rasch Model

The Rasch model quantifies the probability that a given test-taker will answer a given item correctly as a function of two variables: 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 them as a linear function of student ability and item difficulty (Equation 8.1):

$$\log\left(\frac{P_{ni}}{1-P_{ni}}\right) = \theta_n - b_i \tag{8.1}$$

where P_{ni} is the probability that test-taker n will answer item i correctly, 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 (and vice versa). 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 8.2):

$$P_{ni} = \frac{\exp^{\theta_n - b_i}}{1 + \exp^{\theta_n - b_i}} \tag{8.2}$$

The Rasch model makes two assumptions: unidimensionality and local independence. Unidimensionality means that the items on the test measure only a single construct/ability. Chapter 10 shows that the CLT measures three constructs: Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning. Therefore, we fit the Rasch model to the three sections separately. Local independence means that after controlling for the ability, the responses of a student to any two items are uncorrelated. For example, if the answer to an item is implied in a previous item, the residuals of the answers to the two items will be correlated after removing the effect of the latent trait measured by the items. CLT carefully screens items to prevent such clueing.

The fit of the items to the Rasch model can be 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 squared 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 8.3).

$$OutfitMSQ_i = \frac{\sum\limits_n z_{ni}^2}{n} \tag{8.3}$$

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 weighing the squared residuals by the proximity between an item's difficulty and a student's ability (Equation 8.4). For instance, for hard items, prediction errors for high-ability students are weighted more heavily than the prediction errors for low-ability students.

$$InfitMSQ_i = \frac{\sum\limits_n z_{ni}^2 \times Var(X_{ni})}{\sum\limits_n Var(X_{ni})} \tag{8.4}$$

where z_{ni} and $Var(X_{ni})$ are defined as above. The $Var(X_{ni})$ term serves as the item-specific weight because the variance of the response, and thus information gained from a student's response, is maximized when a student's ability matches the item's difficulty (i.e., the probability of answering the item correctly is 0.5). Outfit and infit values have an expectation of 1. Values above 1.5 indicate model misfit whereas values below 0.5 indicate model overfit, meaning that the responses to an item are too predictable and the item does not provide much information. Values between 0.5 and 1.5 are considered productive for measurement (Linacre, 2002). Therefore, we exclude items with an infit or outfit mean square outside this range from the item pool.

ESTIMATING ITEM DIFFICULTIES AND PERSON ABILITIES

IRT models have scale indeterminacy, which means that for any given value of item difficulty, we can find an ability level that retains the same probability of a correct response to that item. That is, without constraining the scale of either the item difficulties or the student abilities, we would have infinitely many ways to describe the relative difference between item difficulty and latent ability, and produce the same response probability. Therefore, the model must be constrained in some way to allow parameter estimation. To solve the issue of scale indeterminacy, one of two approaches is used: 1) identify the scale of test-taker abilities by constraining the latent ability distribution; or 2) identify the scale of item difficulties by constraining the distribution of item parameters. CLT follows the common practice in scale identification by setting the mean of student abilities to 0, which sets a reference point for the estimation of both item difficulties and student abilities.

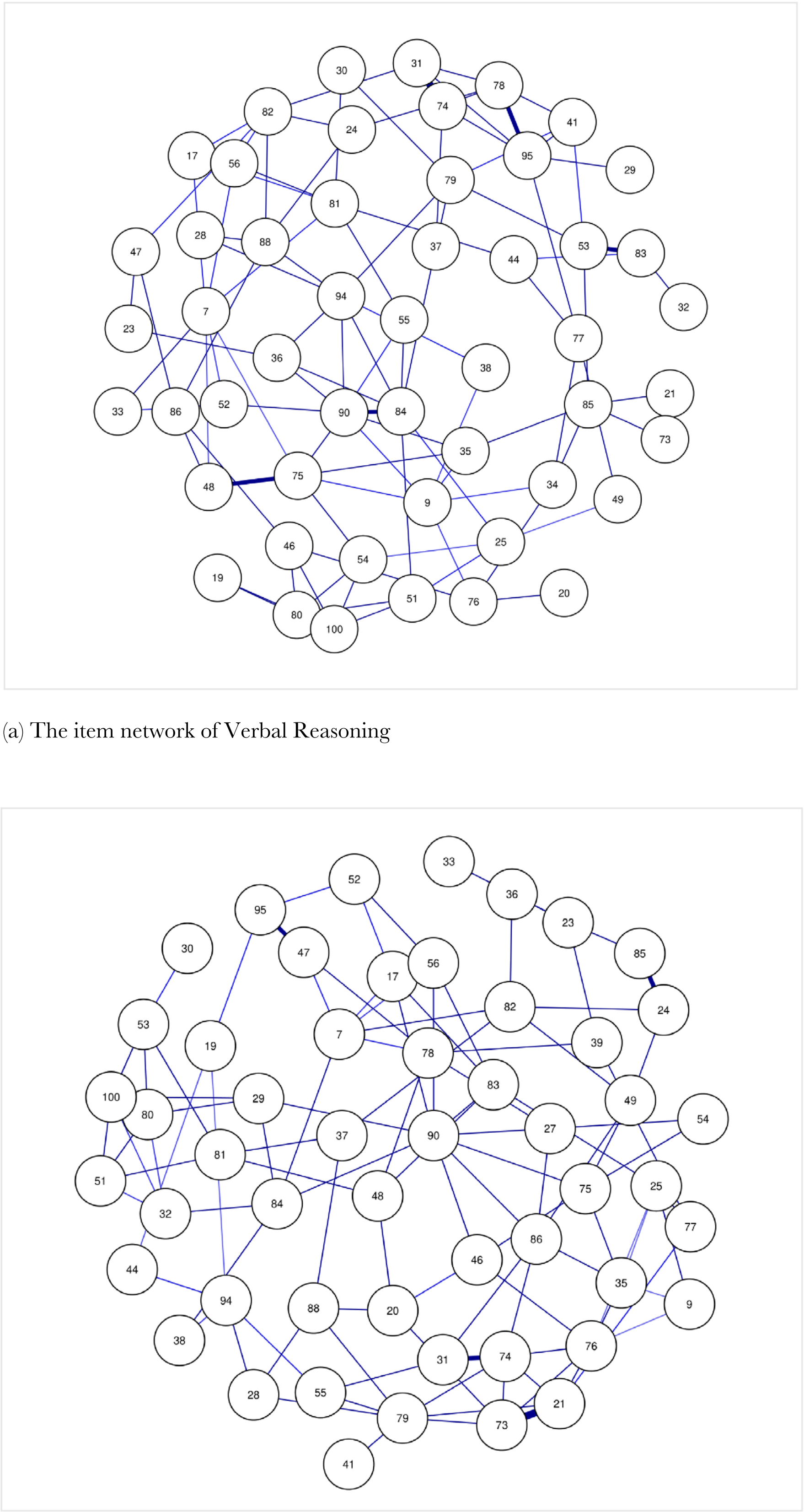
Constraining the mean of student abilities determines the scale for the purposes of parameter estimation, but further steps need to be taken before we can compare estimates obtained from different test forms administered to different groups of students. To illustrate why, suppose that group A takes form 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 identifies test-taker ability, the mean ability distribution will be set to 0 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 have the same difficulty estimate, they cannot be considered equally difficult because the ability level "0" that serves as the reference point in both analyses have 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 a consistent scale and thus comparable. This calibration process is described next. CLT uses the WINSTEPS® software (Linacre, 2023) for Rasch calibrations.

8.3 Fixed Parameter Calibration Using the Common-Item Nonequivalent Groups Design

Items administered to different groups of students can be placed on the same scale when 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 are referred to as anchor items when used to link the scales of two forms. In the Verbal Reasoning and Grammar/Writing sections, the respective anchor items come from common passages. In the Quantitative Reasoning section, the anchor item sets are created from individual items. Anchor items were used to place the item difficulties and the abilities of different groups on the same scale via the fixed-parameter calibration method. First, a base form was selected, which set up an IRT scale for each section. The base form was selected according to the psychometric properties of the overall test form, including reliability and the distribution of item difficulties and point-biserial correlations. To calibrate a new form on the scale of the base form, the difficulty estimates of the anchor items obtained from the base form were constrained to be the same in the new form. This placed the item difficulties as well as the person abilities estimated from the new form onto the scale of the base form. When a new form did not share any items with the base form but had common items with other forms that had already been calibrated and placed on the scale of the base form, those common items were used as anchor items to calibrate the new form. The links between different test forms can be conceptualized as a network with the forms as the nodes and the common items as the edges connecting the nodes. Figure 8.1 presents the network graphs 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).

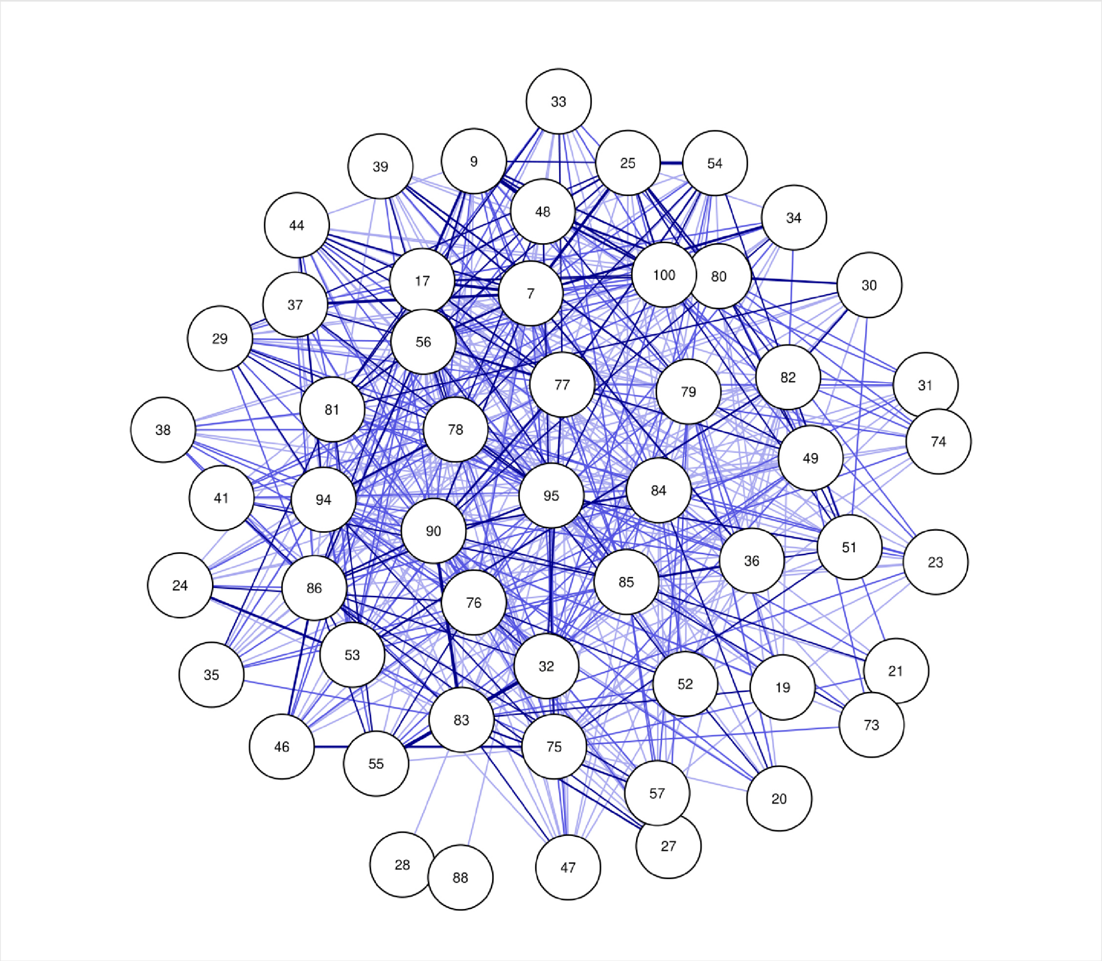
Figure 8.1

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.

CHARACTERISTICS OF ANCHOR ITEMS

The psychometric literature suggests that to achieve a stable calibration, 20% of the items in a form should be anchor items (Kolen & Brennan, 2004). Since there are 40 items in each CLT section, this corresponds to 8 items, a standard which was generally either met or exceeded. In addition, the common items between the forms must remain stable to serve as anchor items. We check anchor item stability to ensure that anchor items function the same in the forms that are linked. Items may function differently across forms due to item drift or mode effects. Item drift refers to the fact that the difficulty of an item may change over time, and 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). To evaluate anchor item stability, we use the displacement statistic from WINSTEPS® (Linacre, 2023). Displacement is the difference between the anchored difficulty of an item and the difficulty estimate that would be obtained if the item was calibrated freely in the new form (i.e., if it was "unanchored"). Items that show a displacement of 0.5 or more logits in absolute value are not used as anchor items (O'Neill et al., 2013). Instead, their difficulty parameters are estimated freely and updated to reflect the most recent estimate for a given mode of administration. Students are scored using the difficulty estimates obtained from the mode in which they were administered the test. Further, we screen the anchor items to ensure that they are high quality items. Specifically, we use item discrimination and (lack of) differential item functioning (DIF) as criteria for selecting anchor items. To evaluate item discrimination, we use the point-biserial correlation, which is defined in Chapter 7. Items with a point-biserial correlation of less than 0.10 and items that show DIF (see Chapter 10 for the quantification of DIF) are excluded from anchor item sets.

DATA CLEANING AND MISSING VALUES

Before the analyses, we apply certain exclusion rules to ensure that the calibration samples are representative of the population, the assumptions of the Rasch model are met, and the parameter estimates are unbiased. First, repeat attempts are excluded from the item calibrations. That is, only the first attempt of each student is used to calibrate the items. Second, we exclude students who did not attempt a given section from the calibrations of that section. Third, missing/blank responses are treated differently in item calibration and scoring; during calibrations, we make a distinction between omitted and not-reached items.

Omitted items are items to which a student did not respond but after which the student continued the test. Given that the student had responses to the subsequent items, we assume that the student saw the omitted items and decided not to respond because they thought that the item was too difficult. In contrast, not-reached items are the missing responses at the end of the test – the missing responses that are not followed by any response. We assume that the student did not actually encounter these items either because they ran out of time or decided to stop the test. 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).

CALIBRATION RESULTS

After these exclusions, we follow the anchoring procedure described in the previous section and fit the Rasch model to each form. The result of this process is a calibrated item bank that has all the difficulty estimates on the same logit scale. Table 8.1 and Table 8.2 show the distribution of the difficulty (b) and the discrimination (r_{pb}) parameters in each section, and Figures 8.2 and 8.3 show the respective frequency distributions. The range of r_{pb} is restricted to ≥ 0.05 because items with a lower point-biserial correlation are excluded from further use and are not displayed in the tables. Items with a point biserial correlation larger than 0.05 but smaller than 0.10 are flagged and reviewed by content experts. As mentioned above, items with an infit or outfit mean square outside the 0.5-1.5 range are also excluded from further use and thus not displayed in the tables.

Table 8.1

Distribution of the Item Difficulties (b)

Section	N	Mean	SD	Min	Max	$b < -2.5$	$-2.5 \leq b \leq 2.5$	$b > 2.5$
VR	1282	-1.04	1.11	-4.45	2.39	9.67%	90.33%	0.00%
GW	1248	-0.94	1.23	-5.48	3.20	11.62%	88.30%	0.08%
QR	1272	-0.18	1.10	-5.54	3.06	2.52%	97.33%	0.16%

Note. The table shows the number of calibrated items and the mean, standard deviation, minimum, and the maximum of the item difficulties. The last column shows the percentage of items with a difficulty value in the [-2.5, 2.5] range. The table only shows items with a point biserial correlation $\geq$ 0.05 and infit and outfit mean squares between 0.5 and 1.5.

Table 8.2

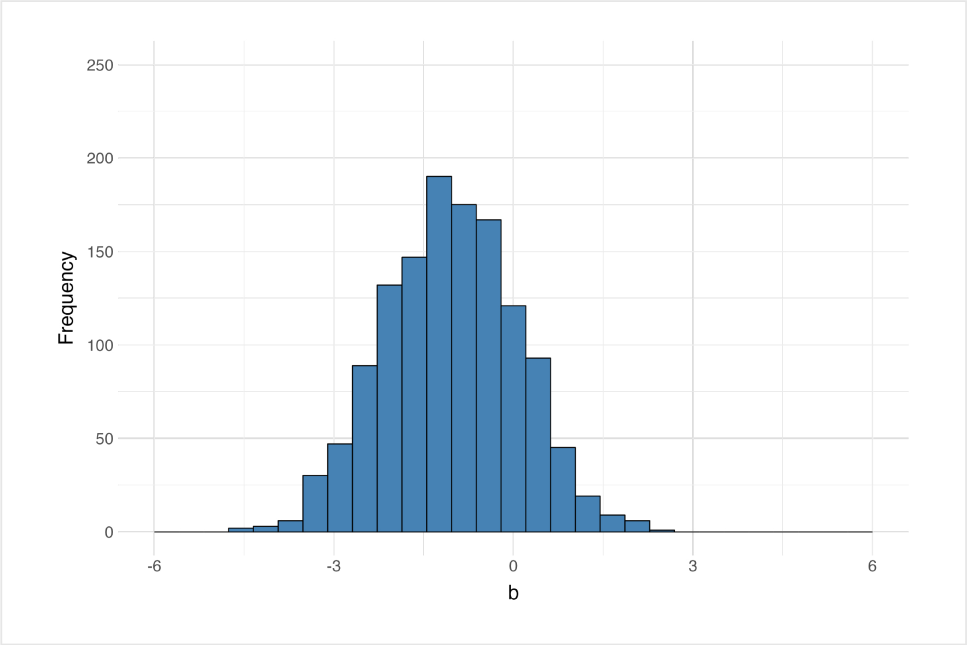
Distribution of the Point-Biserial Correlations (r_{pb})

Section	N	Mean	SD	Min	Max	$0.05 \leq r_{pb} < 0.10$	$0.10 \leq r_{pb} < 0.25$	$r_{pb} \geq 0.25$
VR	1282	0.32	0.10	0.05	0.58	1.25%	23.63%	75.12%
GW	1248	0.31	0.10	0.05	0.63	2.16%	23.32%	74.52%
QR	1272	0.30	0.12	0.05	0.59	4.87%	27.52%	67.61%

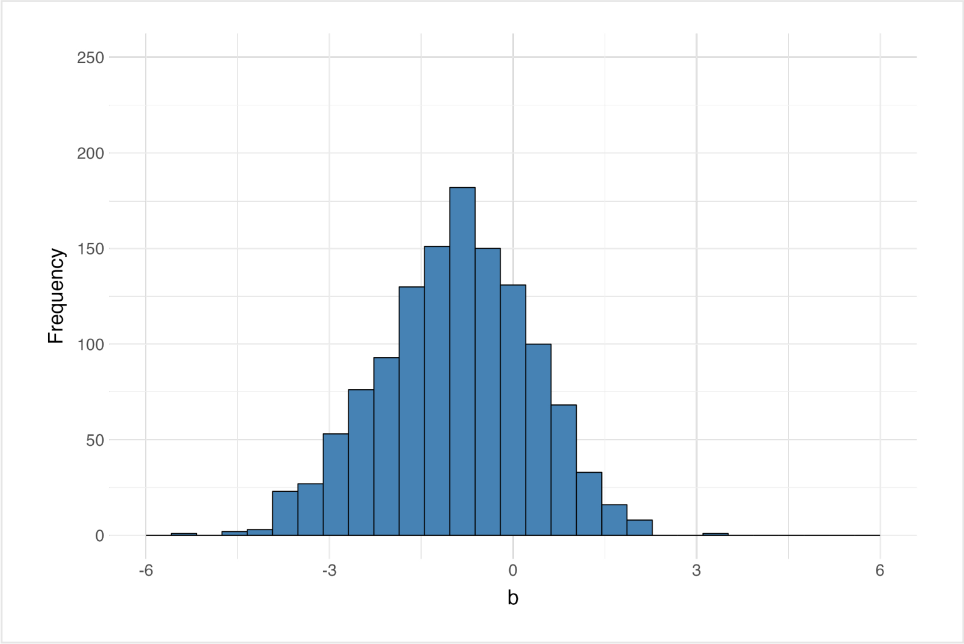
Note. The table shows the number of calibrated items and the mean, standard deviation, minimum, and the maximum of the point-biserial correlations. Also, the last three columns show the percentage of items with a point-biserial correlation in the given range. Items with $r_{pb} < 0.05$ are excluded from operational use and from the tables. The table only shows items with a point biserial correlation ≥ 0.05 and infit and outfit mean squares between 0.5 and 1.5.

Figure 8.2

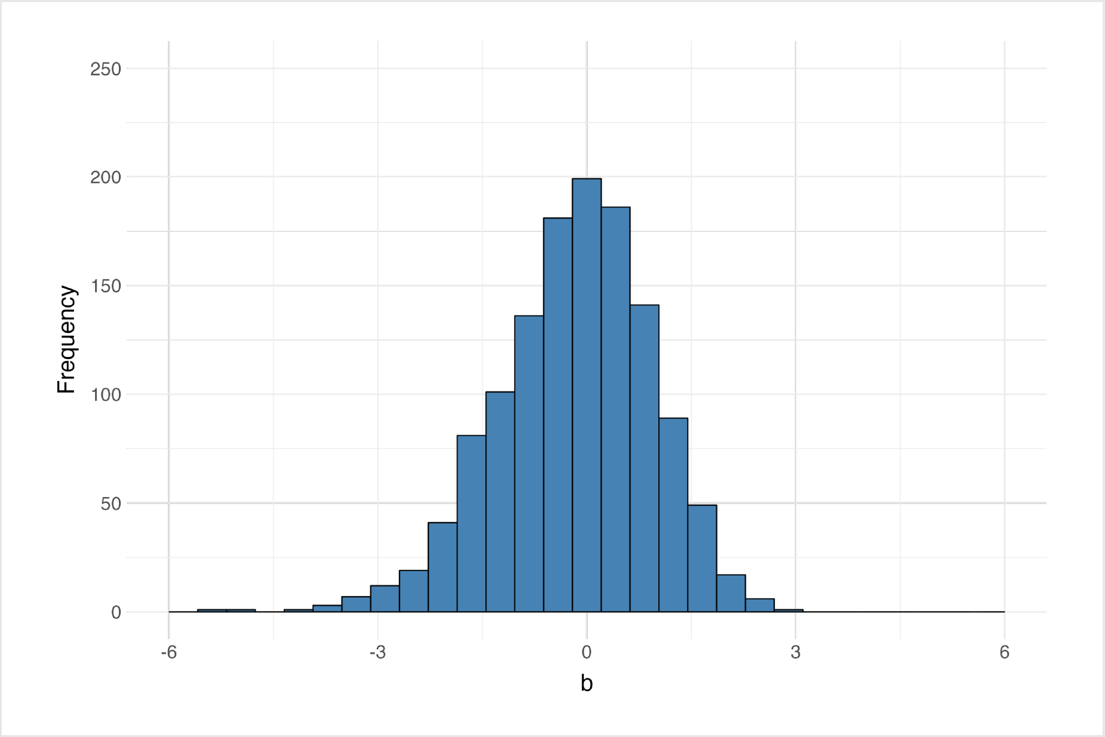
Frequency Distribution of Item Difficulties (*b*) in Each Section



(a) Verbal Reasoning



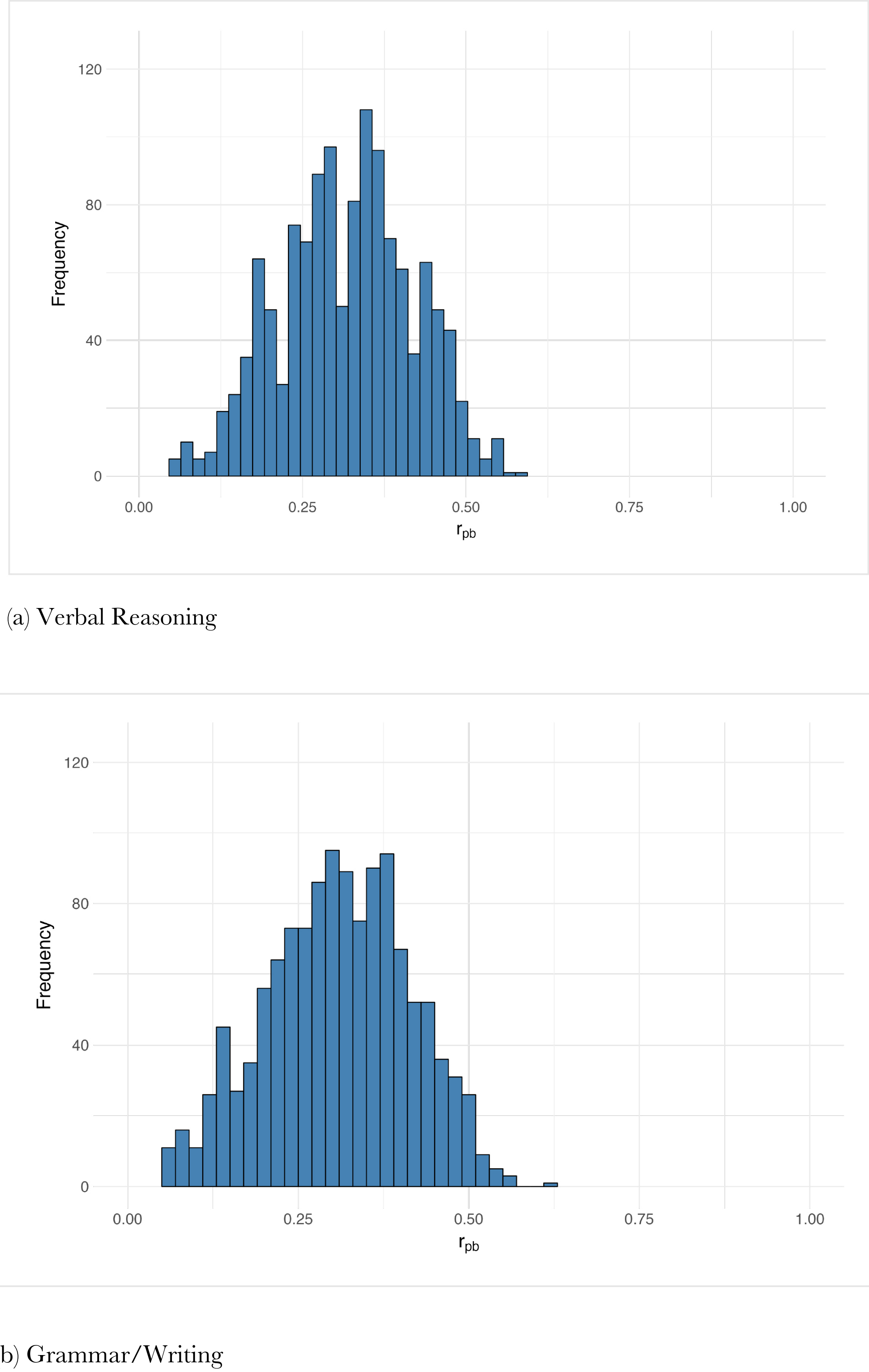
(b) Grammar/Writing



(c) Quantitative Reasoning

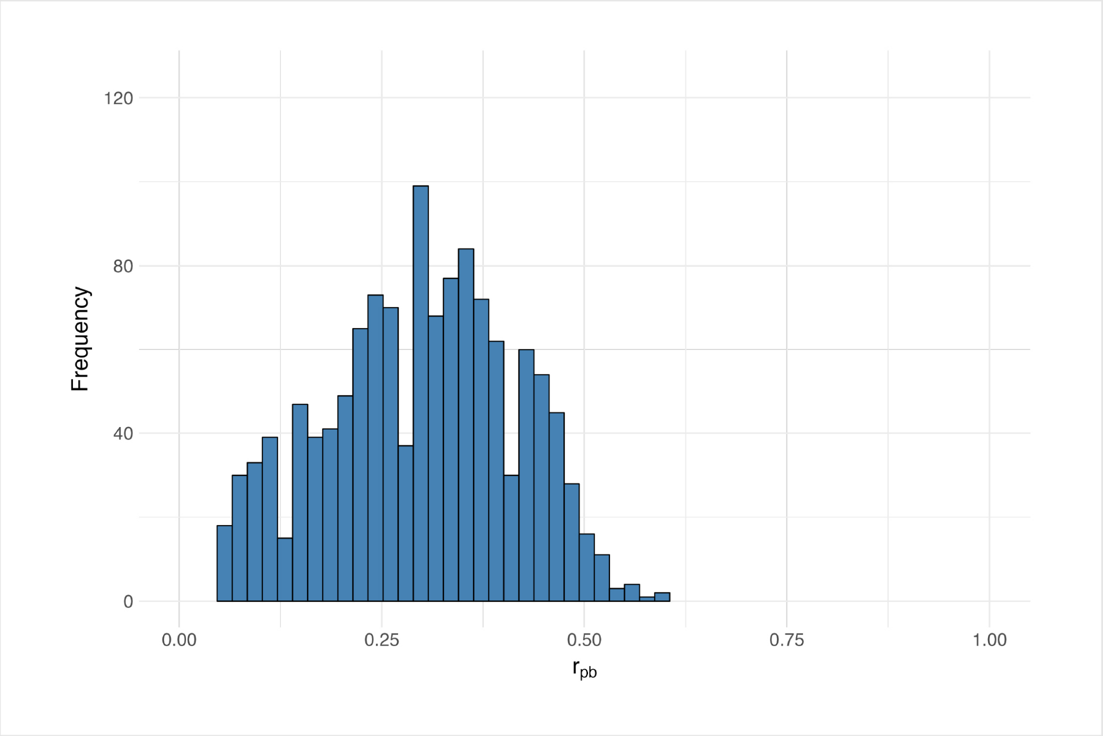
Figure 8.3

Frequency Distribution of Point-Biserial Correlations (r_{pb}) in Each Section



(a) Verbal Reasoning

(b) Grammar/Writing



(c) Quantitative Reasoning

8.4 Scaling and the Reported Scores

Once the items have been calibrated, parallel test forms have been constructed, and the tests have been administered, we take the following steps to calculate the scale score of each student: first, WINSTEPS® (Linacre, 2023) is used to obtain a raw-to-theta conversion table for the test form using the item difficulties. These conversion tables map each raw score on the test to an ability estimate. Next, the ability estimates are used to compute the true score of each student on the base form using its item difficulties. These true scores are the expected scores of each student on each section of the base form, based on their abilities and the difficulties of the items in the base form. The expected score of person n on item i is simply P_{ni} – their probability of answering the item correctly – because it represents the expectation of a binary outcome (correct or incorrect). The total expected score (or the true score) of person n across all the items in a section is the sum of these response probabilities (Equation 8.5).

$$T_n = \sum_i P_{ni} \tag{8.5}$$

true scores are on the same scale. Then, we apply a linear transformation to the true scores to place them on the reported reference CLT scale. The linear transformation has the following form:

$$SS_n = A \times T_n + B \tag{8.6}$$

where SS_n is the scale score of student n, T_n is the true score of student n on the base form, $A = \frac{\sigma(CLT_{SS})}{\sigma(T_{Base})}$, and $B = \mu(CLT_{SS}) - \frac{\sigma(CLT_{SS})}{\sigma(T_{Base})}\mu(T_{Base})$. $\mu(T_{Base})$ and $\sigma(T_{Base})$ are the mean and the standard deviation of the true scores of a representative group of CLT test takers on the base form, and $\mu(CLT_{SS})$ and $\sigma(CLT_{SS})$ are the targeted mean and standard deviation of the scale scores. The values of the slope (A) and the intercept (B) are given in Table 8.3 for each section.

Table 8.3

The Scaling Constants of Each Section

Section	A	B
Verbal Reasoning	0.885	4.836
Grammar/Writing	0.966	1.225
Quantitative Reasoning	0.957	2.291

True scores that are transformed to a scale score lower than the lowest obtainable scale score (LOSS) of 0 or higher than the highest obtainable scale score (HOSS) of 40 are truncated to 0 and 40, respectively. Moreover, perfect raw scores are converted to the HOSS and zero raw scores are converted to the LOSS. 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.

9. RELIABILITY

9.1 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 10.

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). CTT formalizes this concept by separating test scores into two components: a true score and an error component (Equation 9.1):

$$X = T + E \tag{9.1}$$

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) recommends 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 highlight that a test will measure certain ranges of abilities more precisely than others. Therefore, we use CTT to report reliability at the test level and IRT to report measurement precision at different ability levels. Specifically, Sections 9.2-9.4 use Cronbach's alpha (Cronbach, 1951) to estimate reliability and SEM at the test level whereas Section 9.5 calculates CSEM using IRT for each ability level on the logit scale.

9.2 Quantifying Reliability

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

$$r_{XX'} = \frac{s_T^2}{s_X^2} \tag{9.2}$$

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

$$r_{XX'} = 1 - \frac{s_E^2}{s_X^2} \tag{9.3}$$

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:

$$SEM = s_E = \sqrt{s_X^2(1 - r_{XX'})} = s_X\sqrt{(1 - r_{XX'})} \tag{9.4}$$

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 9.6 (Bland & Altman, 1997):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_X^2}\right) \tag{9.5}$$

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).

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 10), we report Cronbach's alpha for each CLT section separately. Cronbach's alpha was calculated using the CTT package (Willse, 2022) in the R programming language (R Core Team, 2023).

9.3 The Reliability and SEM of the Base Form

Table 9.1 presents the Cronbach's alpha and SEM of each section of the base form. Table 9.2 shows the reliabilities for each gender, and Table 9.3 for three ethnicities. The tables show that each section of the CLT base form has high reliability for each analyzed group according to criteria often cited in the psychometric literature (Taber, 2018) as well as criteria used by state scholarships and education departments. For example, Florida's Tax Credit Scholarships require that students take a standardized test with internal consistency/reliability of at least 0.80 (Florida Department of Education, 2023), and Texas Education Agency considers a reliability coefficient between 0.80-0.89 as good (Texas Education Agency, 2022).

Table 9.1

The Reliability of Each Section of the Base Form

Section	α	SEM
Verbal Reasoning	0.87	2.64
Grammar/Writing	0.86	2.58
Quantitative Reasoning	0.87	2.75

Table 9.2

The Reliability of Each Section of the Base Form by Gender

Section	Male		Female	
	α	SEM	α	SEM
Verbal Reasoning	0.88	2.63	0.85	2.65
Grammar/Writing	0.87	2.58	0.84	2.57
Quantitative Reasoning	0.89	2.71	0.82	2.79

Table 9.3

The Reliability of Each Section of the Base Form by Ethnicity

Section	White		African American		Hispanic	
	α	SEM	α	SEM	α	SEM
Verbal Reasoning	0.86	2.63	0.84	2.79	0.86	2.78
Grammar/Writing	0.85	2.56	0.84	2.68	0.87	2.66
Quantitative Reasoning	0.86	2.75	0.86	2.78	0.84	2.79

9.4 Test Information and Conditional Standard Errors of Measurement (CSEM)

The test information function (TIF) computes the amount of information a set of item responses provide about the latent ability parameter. The information provided by an item about the ability parameter depends on the item's difficulty, and is maximized when the item's difficulty matches the student's ability. In the Rasch model, test information is simply the sum of the information provided by individual items (Equation 9.6):

$$I(\theta) = \sum_{i=1}^{k} P_i(1 - P_i) \tag{9.6}$$

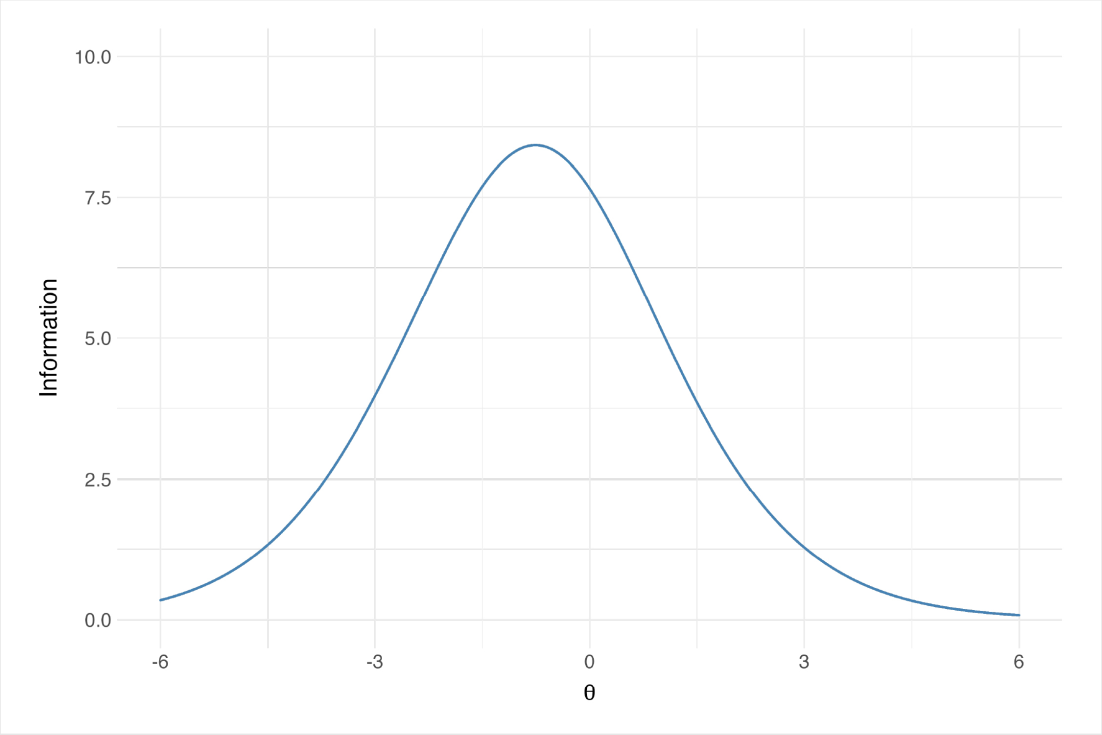
where P_i is the probability of a correct response to item i for a person with ability θ, and k is the total number of items. Test information determines the precision with which a student's ability is estimated by a given set of items. Specifically, test information is inversely related to SEM. Given that test information is a function of the proximity between the test's difficulty and a student's ability, a test will measure different abilities with different levels of precision. While Cronbach's alpha and SEM produce a single estimate of measurement precision for a test form, the IRT framework can estimate CSEM to evaluate measurement imprecision at specific ability levels. Equation 9.7 gives the SEM for a given ability level:

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{9.7}$$

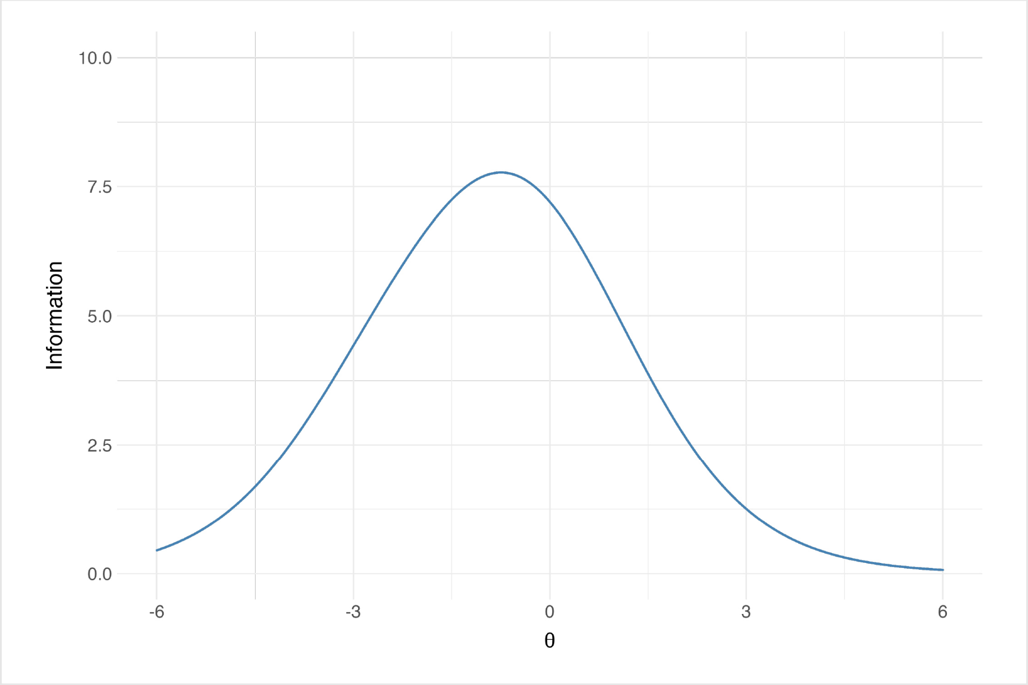
where $I(\theta)$ is the test information function as defined above. Figure 9.1 displays the TIF of the base form for the logit range [-6, 6]. Figure 9.2 shows the CSEM for the same ability range. IRT ability estimates have larger errors at the tails of the distribution than in the middle, which is reflected in Figure 9.2.

Figure 9.1

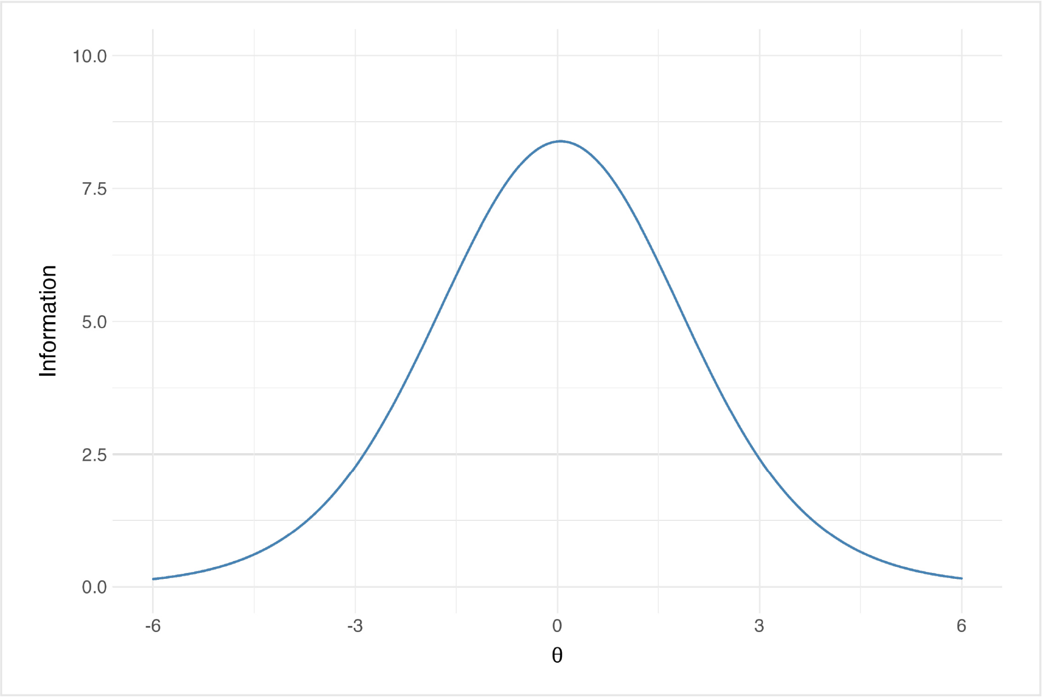
The Test Information Function (TIF) of the Base Form for Each Section



(a) The test information function of the Verbal Reasoning section



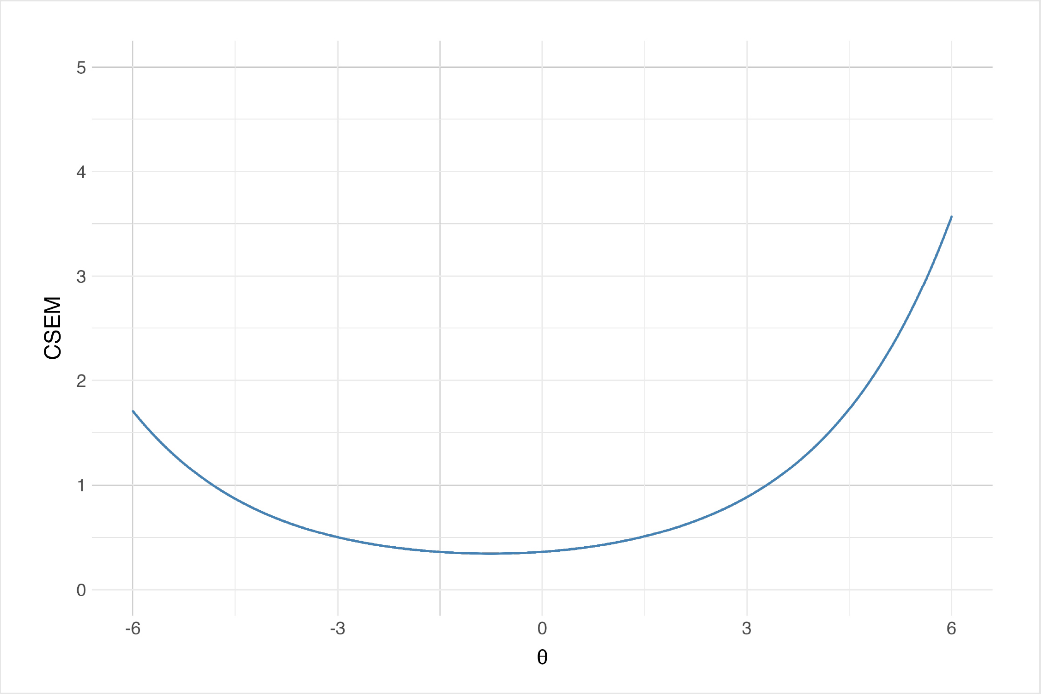
(b) The test information function of the Grammar/Writing section



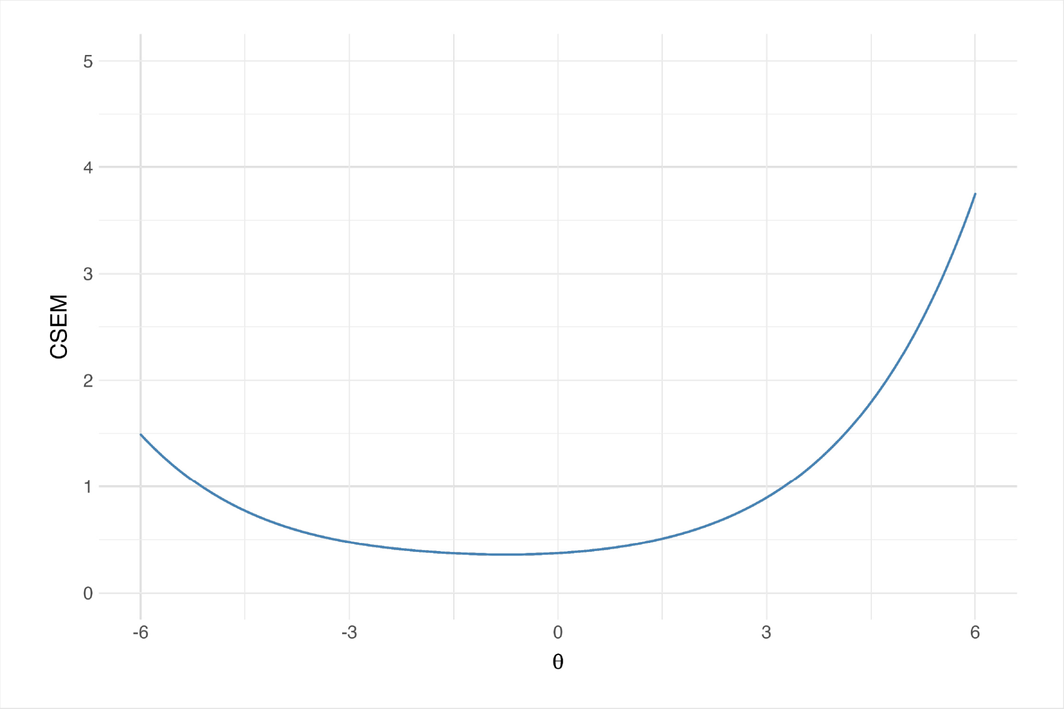
(c) The test information function of the Quantitative Reasoning section

Figure 9.2

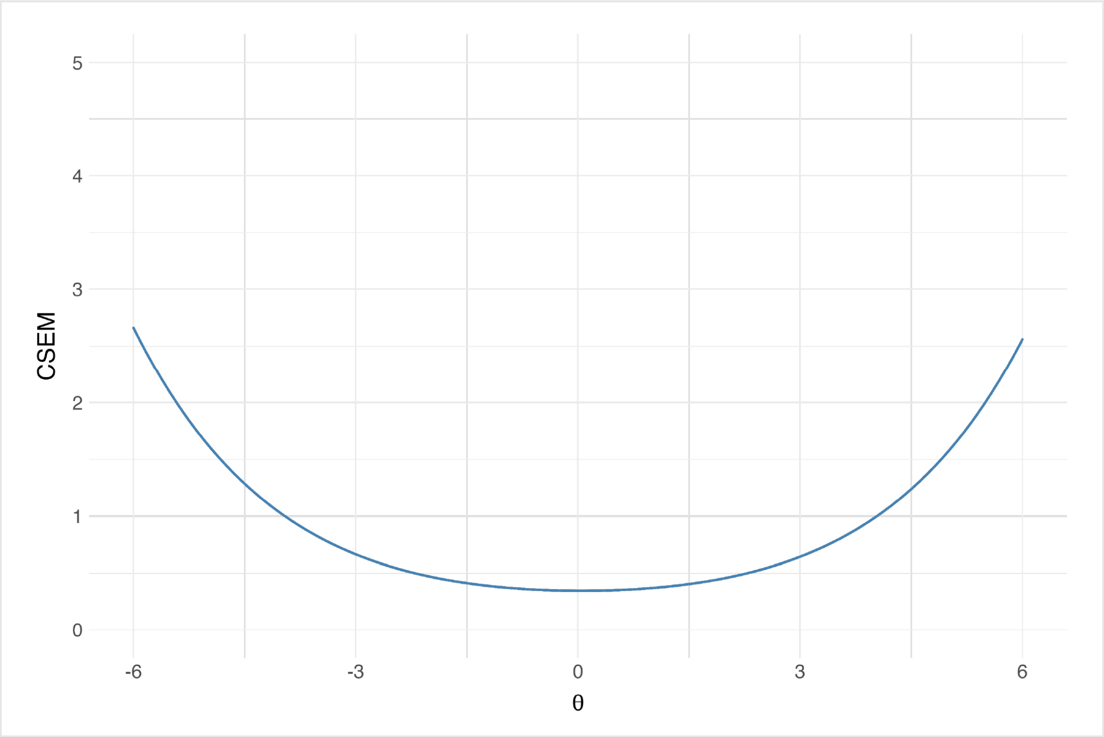
The CSEM of the Base Form for Each Section



(a) The CSEM of the Verbal Reasoning section



(b) The CSEM of the Grammar/Writing section



(c) The CSEM the Quantitative Reasoning section

10. VALIDITY

CHAPTER TEN

10.1 What is Validity?

The Standards for Educational and Psychological Testing defines 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.

10.2 Sources of Validity Evidence

The Standards for Educational and Psychological Testing describes 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 include 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 evidence refers to the degree to which the test scores predict an outcome of interest, and validity generalization refers to the degree to which the test-criterion relationships generalize to new situations. This chapter provides convergent evidence based on the correlations between the CLT and the SAT®.

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

Psychological and cognitive constructs such as student ability, happiness, and creativity 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. Moreover, different factor models can be empirically compared to test if one factor structure fits the data better than others.

Theoretically, three different factor structures could be argued to underlie the CLT. First, it is possible that Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning scores are all manifestations of the same general factor. Alternatively, it may be that Verbal Reasoning and Grammar/Writing represent a common "literacy" factor whereas Quantitative Reasoning represents a "quantitative" factor. However, given that the CLT has three sections designed to measure Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning, it is hypothesized that a three-factor model provides the best description of the test data. Each of these three factors is expected to load on the latent construct underlying each section. To test this hypothesis, CFA was performed on CLT domain scores using the R package lavaan (Rosseel, 2012). The semPlot (Epskamp, 2022) package was used to visualize the factor models.

MODEL FIT

We tested and compared the three different factor structures described in the previous section: a one-factor model where each domain score is an indicator of a single ability, a two-factor model in which the Verbal Reasoning and Grammar/Writing domain scores form one factor while the Quantitative Reasoning domains form a second factor, and a three-factor model in which the Verbal Reasoning domains form one factor, the Grammar/Writing domains form a second factor, and the Quantitative Reasoning domains form a third factor. The fit of each model to the data was examined using exact fit, close fit (MacCallum et al., 1996), and incremental fit indices (Bentler, 1990; Tucker & Lewis, 1973). Exact fit was assessed with the chi-square test, which tests the null hypothesis that the model covariance matrix describes the observed covariance matrix perfectly. A p-value < 0.05 means that the model does not fit the data perfectly. Since the model never fits the data perfectly, we use indices of close fit and incremental fit in addition to exact fit. The close fit indices include the Root Mean Square Error (RMSEA), the RMSEA test of close fit, and the

RMSEA test of not-close fit (MacCallum et al., 1996). RMSEA estimates the degree to which the model deviates from a saturated model (i.e., a model that fits the data perfectly). 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. RMSEA test of close fit tests the null hypothesis that RMSEA is smaller than 0.05, with a p-value < 0.05 indicating that the model does not fit the model very well (but the fit may still be acceptable). RMSEA test of not-close fit tests the null hypothesis that RMSEA is greater than 0.08, with a p-value < 0.05 indicating that the model does not fit the data poorly. The ideal scenario is that the chi-square test of exact fit is non-significant, the RMSEA fit of close fit is also non-significant, but the RMSEA test of not-close fit is significant.

In addition, we used several incremental fit indices. While RMSEA compares the model to a model that fits the data perfectly, incremental fit indices compare it to the baseline model, which only estimates variances and does not model the covariances. In other words, incremental fit indices tell us the degree to which our model is better than the worst possible model. So, the higher they are, the better. We use the Comparative Fit Index (CFI) (Bentler, 1990) and the Tucker-Lewis Index (TLI) (Tucker & Lewis, 1973). Values above 0.90 indicate acceptable fit and values above 0.95 indicate good fit. Further, we report the Goodness of Fit (GFI) and the Adjusted GFI (AGFI), which estimate the proportion of variance in the sample covariance matrix that is explained by the model covariance matrix. Values above 0.90 indicate acceptable fit.

Finally, we used the Akaike Information Criterion (AIC) (Akaike, 1973) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) to compare the models with different numbers of factors. The AIC balances the likelihood of the model against its complexity, with a penalty for the number of parameters estimated by the model. A lower AIC value indicates a model that has a good balance of fit and parsimony. The BIC also considers both the likelihood and the number of parameters, but it penalizes models with more parameters more heavily than the AIC does. Thus, for both AIC and BIC, the model with the lowest value is preferred, with BIC generally favoring simpler models compared to AIC when choosing between models with similar likelihoods. AIC and BIC are given in Equations 10.1 and 10.2:

$$AIC = 2k - 2(L) \tag{10.1}$$

where k is the number of parameters in the model and L is the likelihood of the model given the data, and

$$BIC = k(n) - 2(L) \tag{10.2}$$

where k is the number of parameters, n is the number of observations (the sample size), and L is the likelihood of the model.

THE CORRELATION MATRIX OF THE BASE FORM DOMAIN SCORES

Table 10.1 presents the sample correlation matrix of the domain scores of the base form to provide an overview of its structure. The strongest correlations are between the two domains of the Verbal Reasoning section, Comprehension and Analysis. This is followed by the correlations between Grammar/Writing domains and the Comprehension domain of Verbal Reasoning, and the correlation between Grammar and Writing. The correlations between the Quantitative Reasoning domains are similar in magnitude to the correlations between the Grammar/Writing domains. As expected, they are also larger than the correlations between the Quantitative Reasoning domains and the Verbal Reasoning and Grammar/Writing domains.

Table 10.1

The Correlation Matrix of the Base Form Domain Scores

	Comprehension	Analysis	Grammar	Writing	Geometrical Reasoning	Mathematical Reasoning	Algebra
Comprehension	1						
Analysis	0.763***	1					
Grammar	0.692***	0.675***	1				
Writing	0.716***	0.662***	0.680***	1			
Geometrical Reasoning	0.553***	0.525***	0.527***	0.518***	1		
Mathematical Reasoning	0.505***	0.479***	0.478***	0.484***	0.645***	1	
Algebra	0.576***	0.519***	0.509***	0.550***	0.658***	0.684***	1

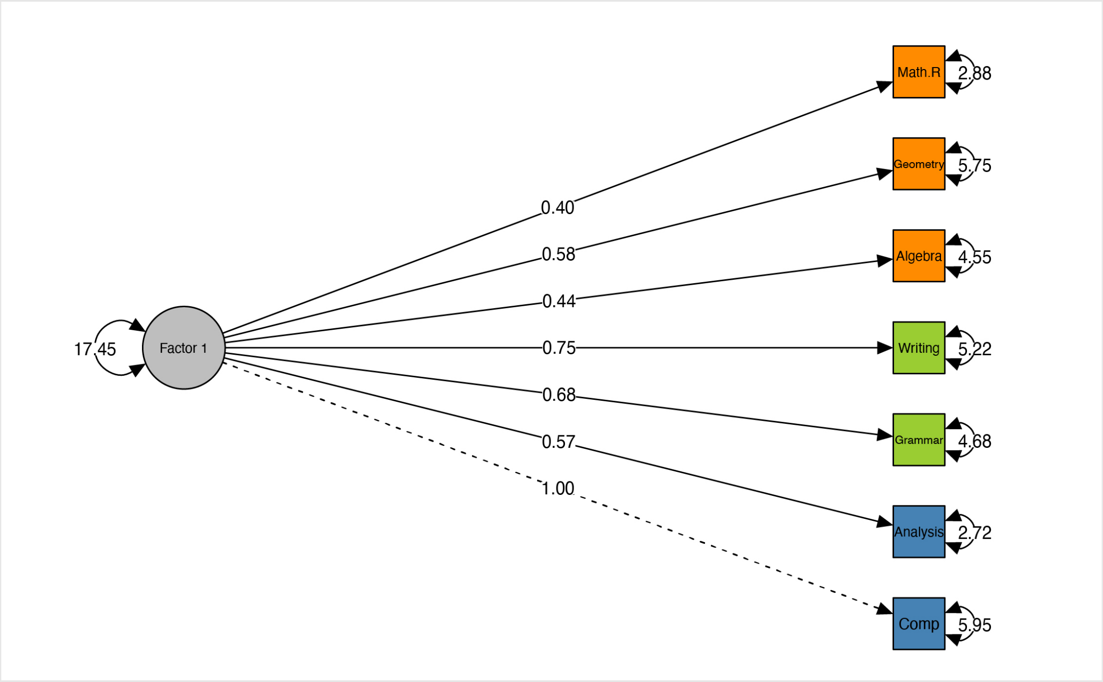
**Correlation is significant at the 0.001 level (2-tailed).

CFA RESULTS

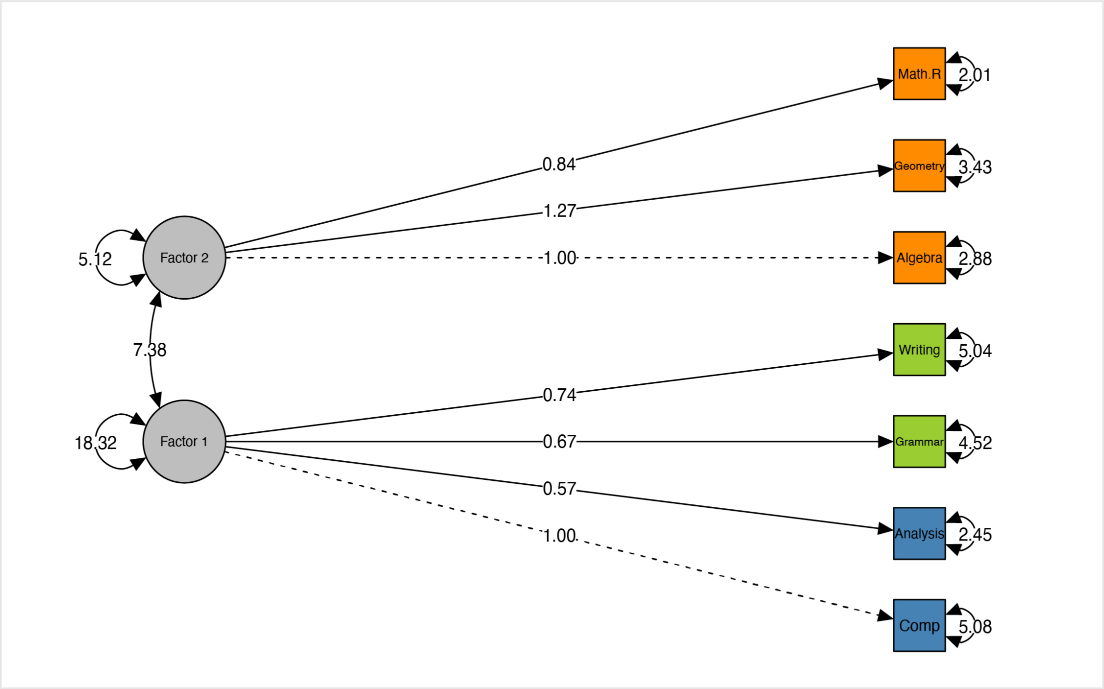
Figure 10.1 visualizes each of the three factor models. The gray circles are the factors and the orange square boxes are the domain scores. Verbal Reasoning domains are colored in blue, Grammar/Writing domains are colored in green, and Quantitative Reasoning domains are colored in orange. The arrows that go from the factors to the domains show the relationship assumed by the model. The fit of these models to the data are evaluated below. The values on these arrows show the factor loadings of each domain. The circular arrows that originate from and end in the same factor show the factor variances, the arrows between the factors show the covariances between the factors, and the arrows that originate from and end in the domains show the variances of the residuals. Model fit is summarized in Table 10.2.

Figure 10.1

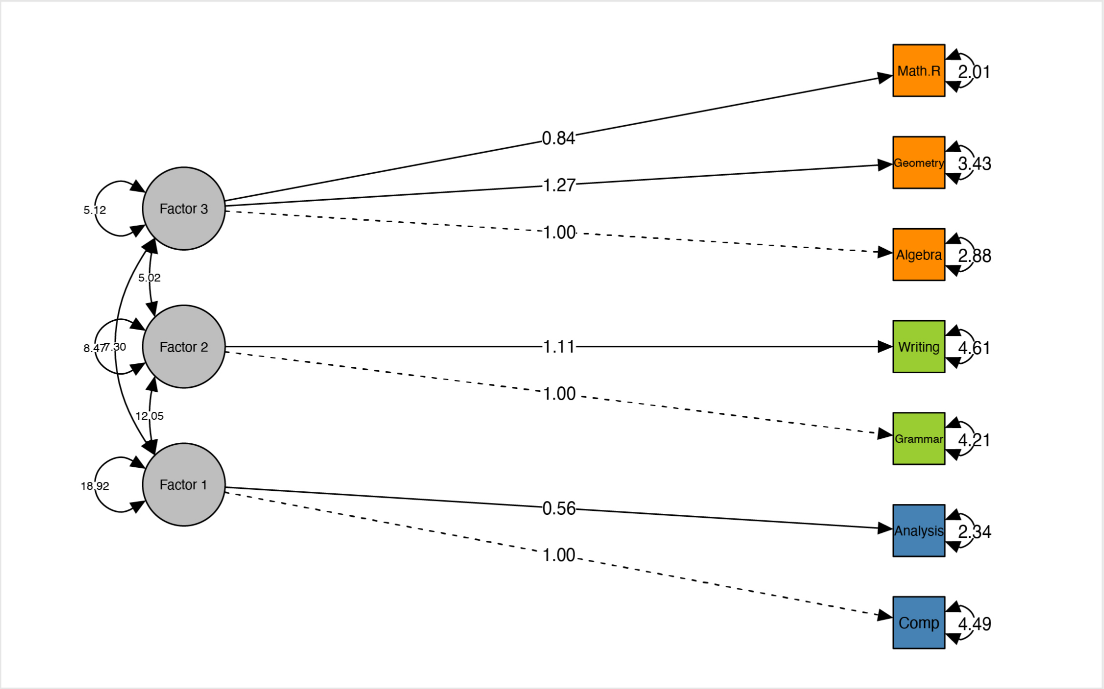
The Structure of the One-Factor, Two-Factor, and Three-Factor CFA Models



(a) The path diagram of the one-factor model



(b) The path diagram of the two-factor model



(c) The path diagram of the three-factor model

Table 10.2

The Fit of the CFA Models

Model	χ^2	df	p	RMSEA	p_{Close}	$p_{Not-Close}$	CFI	TLI	GFI	AGFI	AIC	BIC
1-factor	401.677	14	$< .0001$	0.187	$< .0001$	1.000	0.890	0.836	0.850	0.699	28328.449	28395.144
2-factor	38.929	13	$< .0001$	0.047	0.564	$< .001$	0.993	0.989	0.987	0.973	27932.273	28003.831
3-factor	22.263	11	0.022	0.033	0.891	$< .0001$	0.997	0.995	0.993	0.981	27919.707	28000.693

The results of the one-factor model showed that the model did not fit the data well. The chi-square test was significant ($\chi^2(14) = 401.677, p < 0.0001$), with $RMSEA = 0.187$. The RMSEA test of close fit was also significant ($p < 0.0001$), and the RMSEA test of not-close fit was not significant ($p = 1.000$). Furthermore, $CFI = 0.890$, $TLI = 0.836$, $GFI = 0.850$, and $AGFI = 0.699$ did not reach the acceptable thresholds. For the two-factor model, although the chi-square test of exact fit was significant ($\chi^2(13) = 38.929, p < 0.0001$), all the other fit indices indicated good fit. The RMSEA was 0.047, test of close fit was not significant ($p = 0.564$), and the test of not-close fit was significant ($p = 0.001$). Moreover, $CFI = 0.993$, $TLI = 0.989$, $GFI = 0.987$, and $AGFI = 0.973$ were all very high. Similarly, for the three-factor model the chi-square test of exact fit was significant ($\chi^2(11) = 22.263$), $p = 0.022$), but all the other fit indices showed very good fit. The RMSEA was 0.033, test of close fit was not significant ($p = 0.891$), and the test of not-close fit was significant ($p < 0.0001$). $CFI = 0.997$, $TLI = 0.995$, $GFI = 0.993$, and $AGFI = 0.981$ were all very high.

As would be expected from these results, model comparisons favored the two-factor model ($AIC = 27932.273, BIC = 28003.831$) and the three-factor model ($AIC = 27919.707, BIC = 28000.693$) over the one-factor model ($AIC = 28328.449, BIC = 28395.144$). Moreover, AIC favored the three-factor model over the two-factor model, with a difference of $27932.373 - 27919.707 = 12.666$. The BIC ($28003.831 - 28000.693 = 3.138$) also supported the three-factor model in comparison to the two-factor model, but to a smaller degree than the AIC. Burnham and Anderson (2004) suggested that an AIC difference of more than 10 provides significant support for the model with the lower AIC. Thus, these findings indicate that the CLT measures three related but distinct constructs represented by the Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning sections.

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

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 such as raw scores or IRT ability estimates (as in the analyses reported below). 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 10.3) (Zwick, 2012):

$$\hat{\alpha}_{MH} = \frac{\sum_k N_{R1k} \times N_{F0k}}{\sum_k N_{R0k} \times N_{F1k}} \tag{10.3}$$

where NR1k is the number of test takers in the reference group who answered correctly at the k-th ability level, NF1k is the number of test takers in the focal group who answered correctly at the k-th ability level, NR0k is the number of test takers in the reference group who answered incorrectly at the k-th ability level, and NF0k is the number of test takers in the focal group who answered incorrectly at the k-th ability level. As a measure of effect size, Holland and Thyer (1985) developed the MH index to express MH on the Educational Testing Services (ETS) delta scale (Zwick, 2012):

$$\Delta_{MH} = -2.35 \ln \hat{\alpha}_{MH} \tag{10.4}$$

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 ΔMH (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):

$$MH_{\chi^2} = \frac{(\mid \sum_k N_{R1k} - \sum_k E(N_{R1k}) \mid - \frac{1}{2})^2}{\sum_k Var(N_{R1k})} \qquad (10.5)$$

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 10.6 (Zwick, 2012):

$$\frac{|\Delta_{MH}| - 1}{SE_{\Delta_{MH}}} > 1.645 \qquad (10.6)$$

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 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.

We analyzed DIF for two demographic variables: ethnicity and gender. 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. The MH statistics were calculated using the difR R package (Magis et al., 2010). Tables 10.3-10.5 show the distribution of the base form items across the ETS categories, grouped by gender and ethnicity. Table 10.3 compares male students to female students, Table 10.4 compares White students to African American students, and Table 10.5 compares White students to Hispanic/Latino students. The tables show that the vast majority of the items had no DIF (category A). Moreover, comparisons of females with males and Hispanic/Latino students with White students showed no items in the C category. The comparisons of African-American students with White students showed a total of five items in the C category across the three sections, with four of them appearing to favor African-American students. In general, items that appeared to favor one group were accompanied by items that appeared to favor the other group, suggesting that no group was exclusively favored over the others. Items that show DIF are reviewed by content experts to ensure that they do not lead to bias.

Table 10.3

DIF Results for Males and Females

Section	Total	A	B+	B-	C+	C-
Verbal Reasoning	40	37	2	1	0	0
Grammar/Writing	40	35	3	2	0	0
Quantitative Reasoning	40	37	2	1	0	0

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

Table 10.4

DIF Results for White and African-American Students

Section	Total	A	B+	B-	C+	C-
Verbal Reasoning	40	37	1	2	0	0
Grammar/Writing	40	37	2	0	1	0
Quantitative Reasoning	40	35	0	1	3	1

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

Table 10.5

DIF Results for White and Hispanic Students

Section	Total	A	B+	B-	C+	C-
Verbal Reasoning	40	40	0	0	0	0
Grammar/Writing	40	38	2	0	0	0
Quantitative Reasoning	40	39	0	1	0	0

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

10.5 Validity Based on 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, 2023). 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 10.6. 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 10.6

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

11. NORMING

This chapter provides national norms for CLT scores through the CLT-SAT® concordance study conducted in spring 2023 (Classic Learning Initiatives, 2023). 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 10 for the correlations) (Dorans, 2004; Dorans & Walker, 2007). Furthermore, Chapter 9 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 derive 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

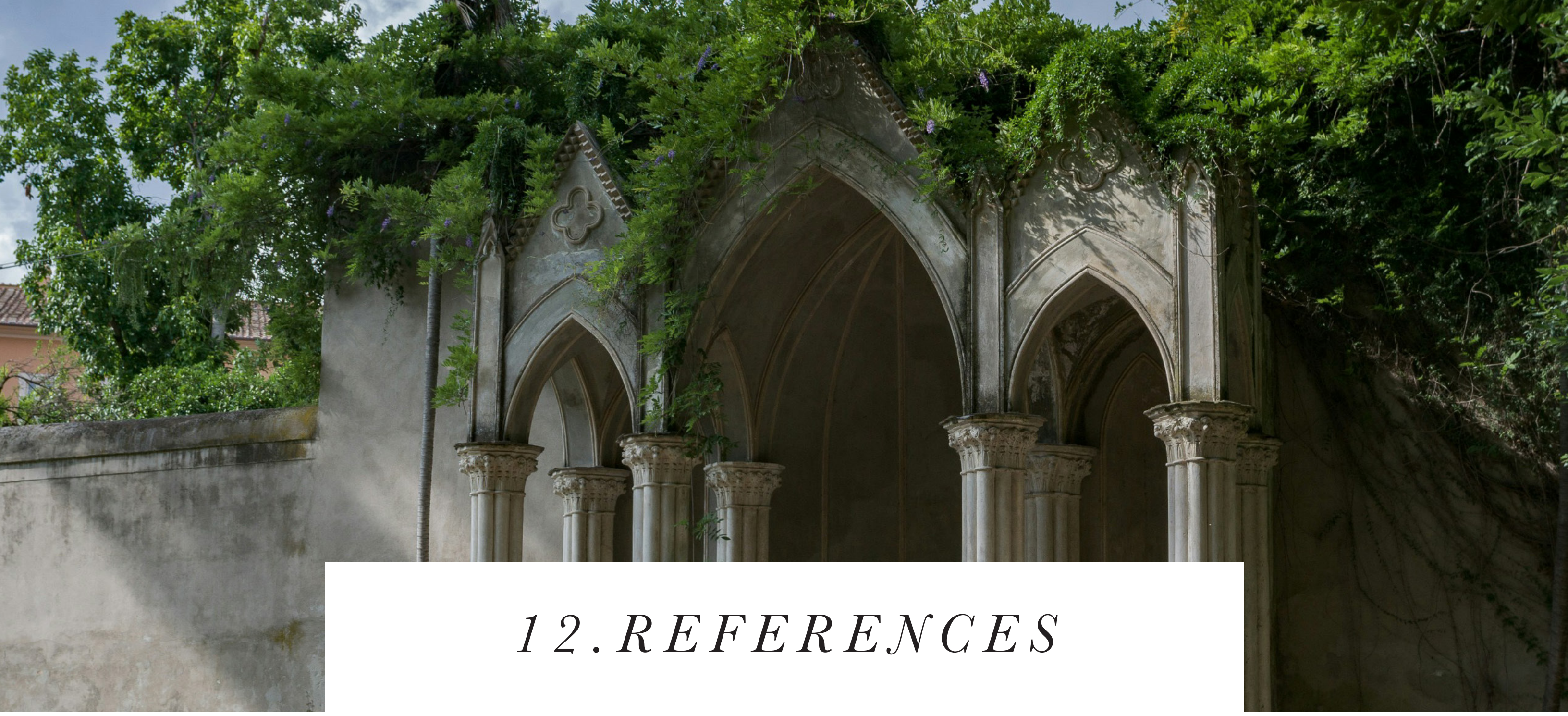
The Concordance Between CLT Verbal Reasoning + Grammar/Writing and SAT EBRW Scores, and the Corresponding SAT National Percentiles

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

CHAPTER TWELVE

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, B.N. and Csaki, F., Eds., International Symposium on Information Theory, 267-281.

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.

Becker B, Debeer D, Sachse K, Weirich S (2021). "Automated Test Assembly in R: The eatATA Package." Psych, 3(2), 96–112.https://www.mdpi.com/2624-8611/3/2/10.

Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. Psychological Bulletin, 107, 238-246.http://dx.doi.org/10.1037/0033-2909.107.2.238

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

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research, 33(2), 261-304. https://doi.org/10.1177/0049124104268644

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. (2023). 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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/bf02310555

Crocker, L. M., & Algina, J. (2008). Introduction to classical and modern test theory. Cengage learning.

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 standardization1,2. *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 (Statistics for Social and Behavioral Sciences). Springer, New York, NY.

Epskamp S (2022). semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output. R package version 1.1.6, https://CRAN.R-project.org/package=semPlot/

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

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). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the mantel-haenszel procedure. ETS Research Report Series, 1986(2). https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

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), p.878 https://doi.org/10.1177/0013164499594004

Linacre, J. M. (2023). Winsteps® Rasch measurement computer program (Version 5.6.0). Portland, Oregon: Winsteps.com

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power Analysis and Determination of Sample Size for Covariance Structure Modeling. Psychological Methods, 1, 130-149.https://doi.org/10.1037/1082-989X.1.2.130

Magis D, Beland 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.

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.

O'Neill, T., Peabody, M., Tan, R. J. B., & Du, Y. (2013). How much item drift is too much? Rasch Measurement Transactions, 27(3), 1423-1424.

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 (Statistics for Social and Behavioral Sciences). Springer, New York, NY.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.https://www.R-project.org/.

Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D (2012). "qgraph: Network Visualizations of Relationships in Psychometric Data." *Journal of Statistical Software*, 48(4), 1–18.

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6(2), 461–464. http://www.jstor.org/stable/2958889

Taber, K. S. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Res Sci Educ 48, 1273–1296 (2018). 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). Retrieved from 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, 1-10. https://doi.org/10.1007/BF02291170

Willse. J. T. (2018). CTT: Classical Test Theory Functions. R package version 2.3.3, https://CRAN.R-project.org/package=CTT.

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

Zieky, M. (2003). A DIF primer. Educational Testing Service. Retrieved from 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